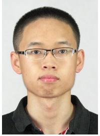
Statistical and Algorithmic Foundations of Reinforcement Learning

Yuting Wei

Statistics & Data Science, Wharton
University of Pennsylvania

PKU, July 2023

Our wonderful collaborators



Gen Li

UPenn → CUHK

Shicong Cen

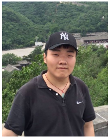
CMU

Chen Cheng

Stanford

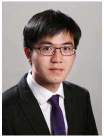
Laixi Shi

CMU → Caltech



Yuling Yan

Princeton → MIT

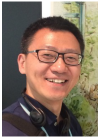


Changxiao Cai

UPenn → UMich

Wenhao Zhan

Princeton



Yuantao Gu

Tsinghua

Jason Lee

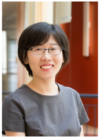
Princeton

Jianqing Fan

Princeton

Yuxin Chen

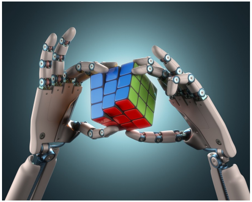
UPenn

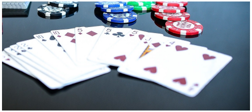


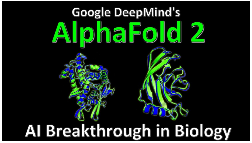
Yuejie Chi

CMU

Recent successes in reinforcement learning (RL)







RL holds great promise in the next era of artificial intelligence.

Recap: Supervised learning

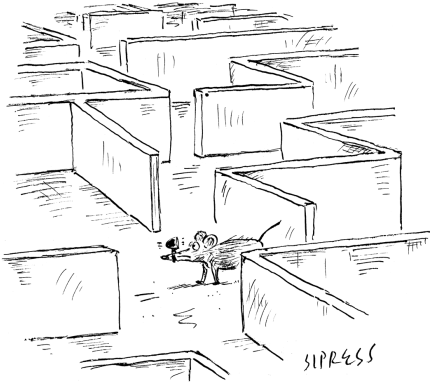
Given i.i.d training data, the goal is to make prediction on unseen data:

— *pic from internet*

Reinforcement learning (RL)

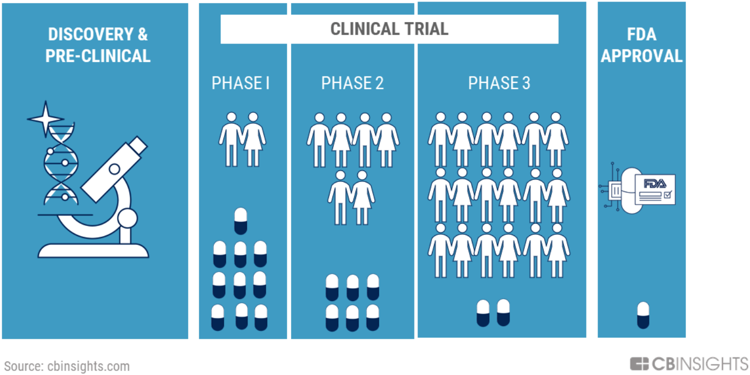
In RL, an agent learns by interacting with an environment.

- no training data
- trial-and-error
- maximize total rewards
- delayed reward



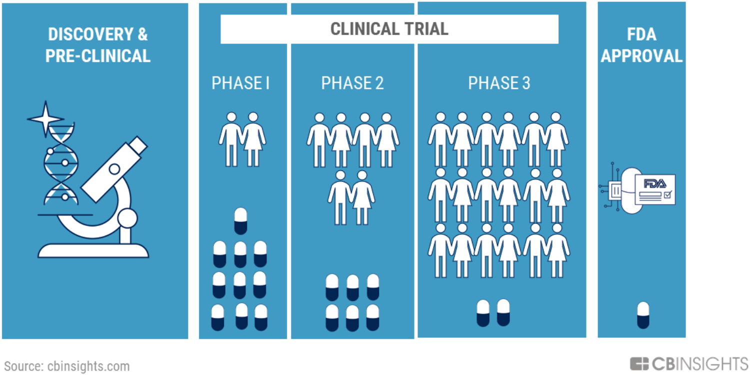
"Recalculating ... recalculating ..."

Sample efficiency



- prohibitively large state & action space
- collecting data samples can be expensive or time-consuming

Sample efficiency



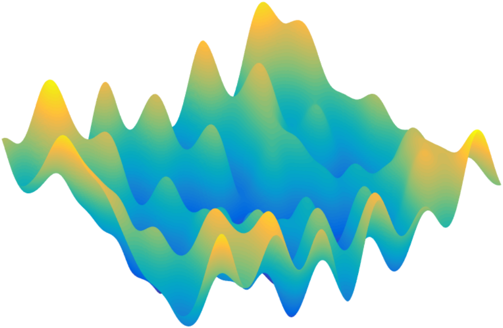
- prohibitively large state & action space
- collecting data samples can be expensive or time-consuming

Challenge: design sample-efficient RL algorithms

Computational efficiency

Running RL algorithms might take a long time . . .

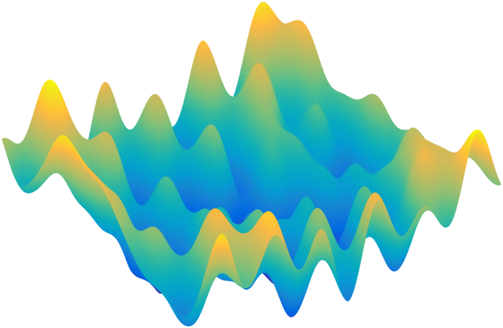
- enormous state-action space
- nonconvexity



Computational efficiency

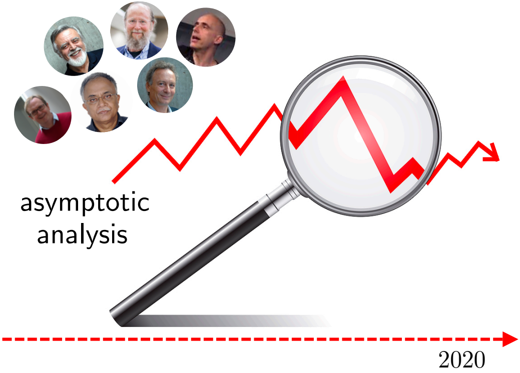
Running RL algorithms might take a long time . . .

- enormous state-action space
- nonconvexity



Challenge: design computationally efficient RL algorithms

Theoretical foundation of RL



Statistical Science
1986, Vol. 1, No. 2, 276–284

The Contributions of Herbert Robbins to Mathematical Statistics

Tze Leung Lai and David Siegmund

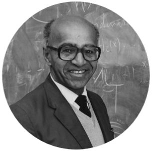
2. STOCHASTIC APPROXIMATION AND ADAPTIVE DESIGN

In 1951, Robbins and his student, Sutton Monro, founded the subject of stochastic approximation with the publication of their celebrated paper [26]. Consider the problem of finding the root θ (assumed unique) of an equation $g(x) = 0$. In the classical

4. SEQUENTIAL EXPERIMENTATION AND OPTIMAL STOPPING

The well known "multiarmed bandit problem" in the statistics and engineering literature, which is prototypical of a wide variety of adaptive control and design problems, was first formulated and studied by Robbins [28]. Let A, B denote two statistical populations with finite means μ_A, μ_B. How should we draw a

Herbert Robbins



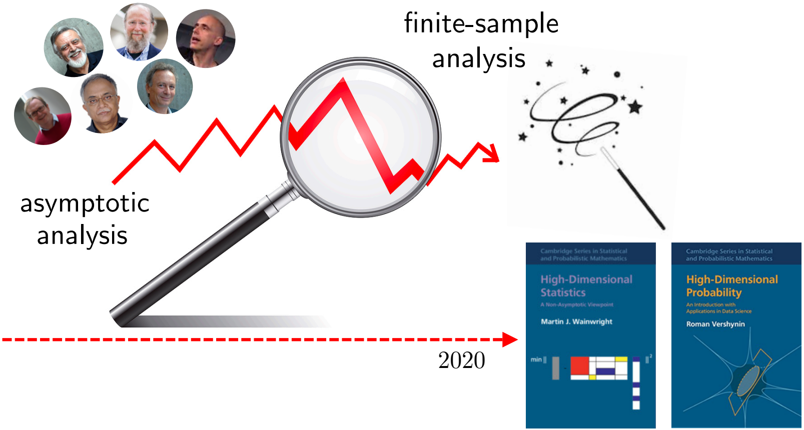
David Blackwell

David Blackwell, 1919–2010: An explorer in mathematics and statistics

Peter J. Bickel[a,1]

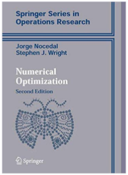
Blackwell channel. He also began to work in dynamic programming, which is now called reinforcement learning. In a series of papers, Blackwell gave a rigorous foundation to the theory of dynamic programming, introducing what have become known as Blackwell optimal policies.

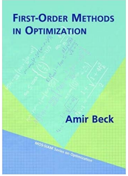
Theoretical foundation of RL

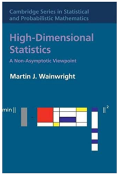


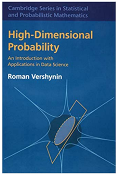
Understanding sample efficiency of RL requires a modern suite of non-asymptotic analysis tools

This tutorial





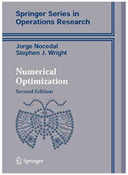


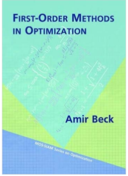
(large-scale) optimization

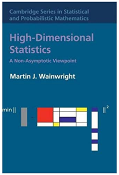
(high-dimensional) statistics

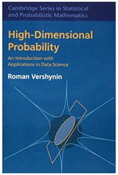
Demystify sample- and computational efficiency of RL algorithms

This tutorial







(large-scale) optimization (high-dimensional) statistics

Demystify sample- and computational efficiency of RL algorithms

Part 1. **basics, model-based and model-free RL**

Part 2. **robust RL, offline RL and multi-agent RL**

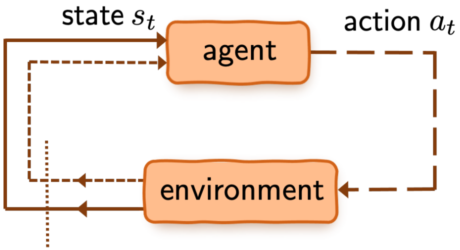
Part 3. **policy optimization**

Outline (Part 1)

- Basics: Markov decision processes
- Basic dynamic programming algorithms
- Model-based RL ("plug-in" approach)
- Value-based RL (a model-free approach)

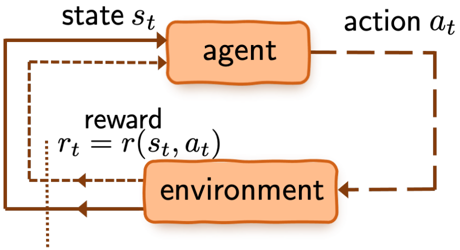
Basics: Markov decision processes

Markov decision process (MDP)



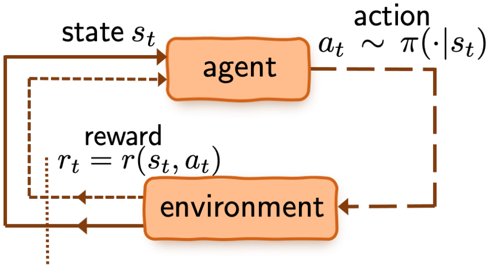
- $\mathcal{S}$: state space
- $\mathcal{A}$: action space

Markov decision process (MDP)



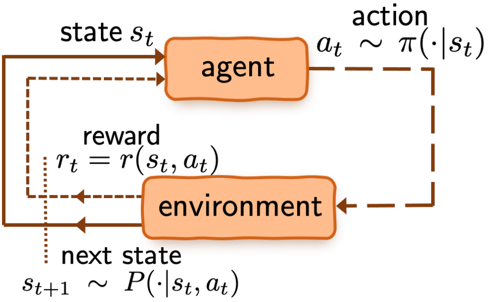
- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s, a) \in [0, 1]$: immediate reward

Infinite-horizon Markov decision process



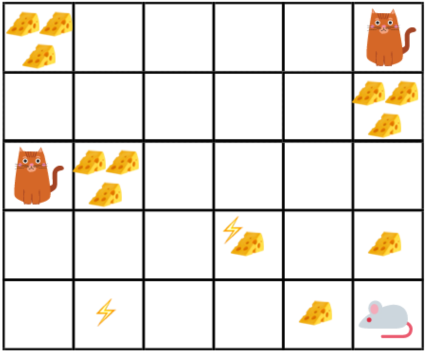
- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s,a) \in [0,1]$: immediate reward
- $\pi(\cdot|s)$: policy (or action selection rule)

Infinite-horizon Markov decision process

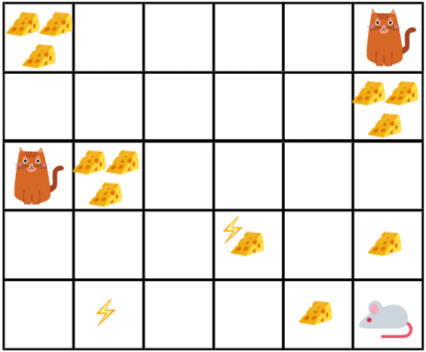


- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s,a) \in [0,1]$: immediate reward
- $\pi(\cdot|s)$: policy (or action selection rule)
- $P(\cdot|s,a)$: unknown transition probabilities

Help the mouse!

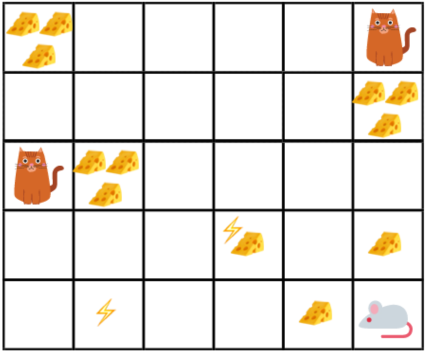


Help the mouse!



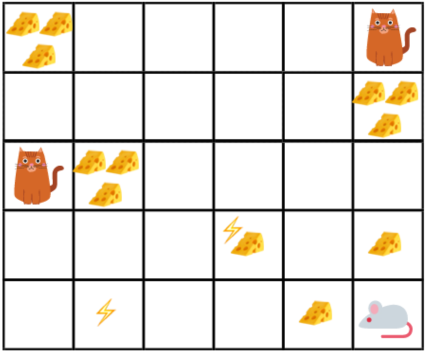
- state space $\mathcal{S}$: positions in the maze

Help the mouse!



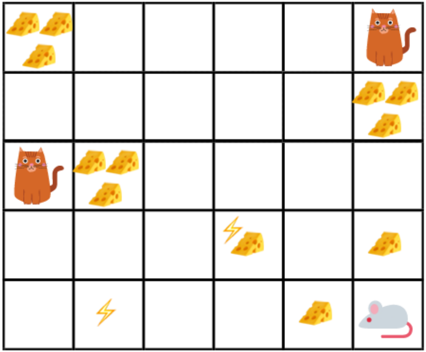
- state space $\mathcal{S}$: positions in the maze
- action space $\mathcal{A}$: up, down, left, right

Help the mouse!



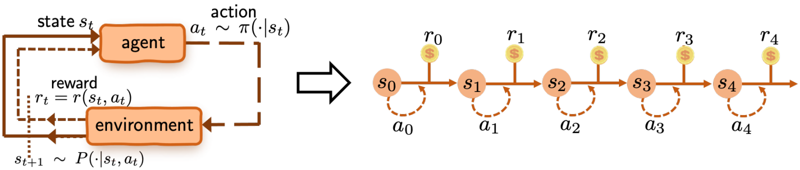
- state space $\mathcal{S}$: positions in the maze
- action space $\mathcal{A}$: up, down, left, right
- immediate reward r: cheese, electricity shocks, cats

Help the mouse!



- state space $\mathcal{S}$: positions in the maze
- action space $\mathcal{A}$: up, down, left, right
- immediate reward r: cheese, electricity shocks, cats
- policy $\pi(\cdot|s)$: the way to find cheese

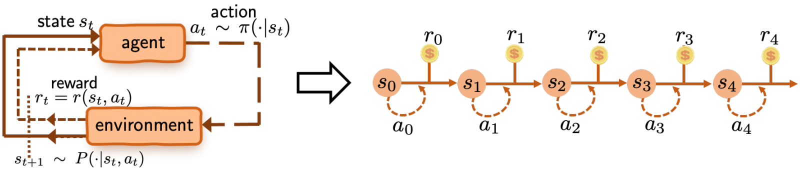
Value function



Value of policy π: cumulative discounted reward

$$\forall s \in \mathcal{S}: \quad V^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s\right]$$

Value function



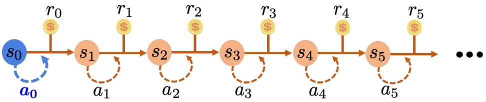
Value of policy π: cumulative discounted reward

$$\forall s \in \mathcal{S}: \quad V^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s\right]$$

- $\gamma \in [0, 1)$: discount factor
  - take $\gamma \to 1$ to approximate long-horizon MDPs
  - **effective horizon**: $\frac{1}{1-\gamma}$

Q-function (action-value function)

$Q^\pi(s_0, a_0)$

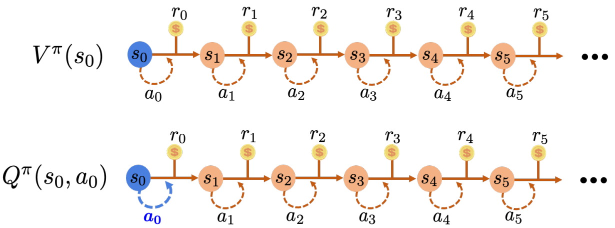


Q-function of policy π:

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A}: \quad Q^\pi(s, a) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, s_0 = s, a_0 = a\right]$$

- $(\require{cancel}\cancel{a_0,}\ s_1, a_1, s_2, a_2, \cdots)$: induced by policy π

Q-function (action-value function)

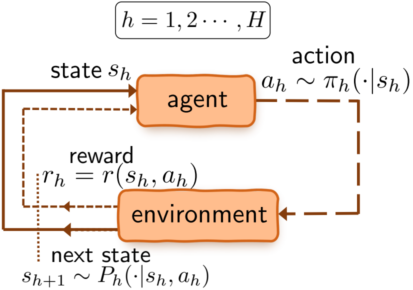


Q-function of policy π:

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A}: \quad Q^\pi(s, a) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, s_0 = s, a_0 = a\right]$$

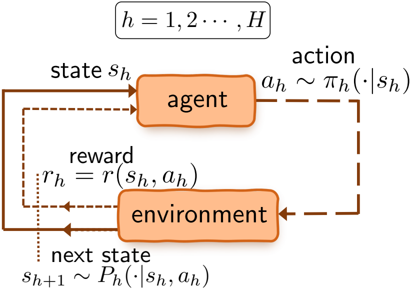
- $(\cancel{a_0}, s_1, a_1, s_2, a_2, \cdots)$: induced by policy π

Finite-horizon MDPs



- H: horizon length
- $\mathcal{S}$: state space with size S
- $\mathcal{A}$: action space with size A
- $r_h(s_h, a_h) \in [0, 1]$: immediate reward in step h
- $\pi = \{\pi_h\}_{h=1}^{H}$: policy (or action selection rule)
- $P_h(\cdot \mid s, a)$: transition probabilities in step h

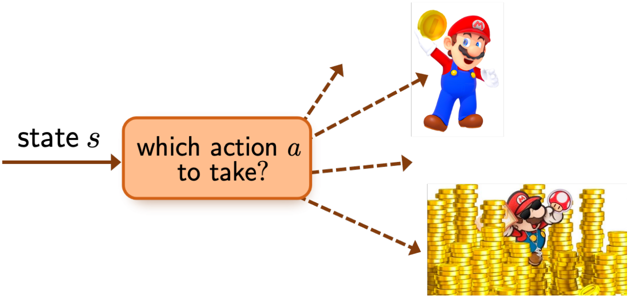
Finite-horizon MDPs



value function: $V_h^\pi(s) := \mathbb{E}\left[\sum_{t=h}^{H} r_h(s_h, a_h) \,\Big|\, s_h = s\right]$

Q-function: $Q_h^\pi(s, a) := \mathbb{E}\left[\sum_{t=h}^{H} r_h(s_h, a_h) \,\Big|\, s_h = s, a_h = a\right]$

Optimal policy and optimal value



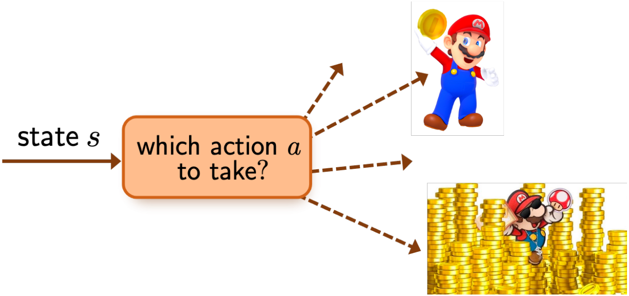
optimal policy $\pi^\star$: maximizing value function $\max_\pi V^\pi$

Proposition (Puterman'94)

For infinite horizon discounted MDP, there always exists a deterministic policy $\pi^\star$*, such that*

$$V^{\pi^\star}(s) \geq V^\pi(s), \quad \forall s, \text{ and } \pi.$$

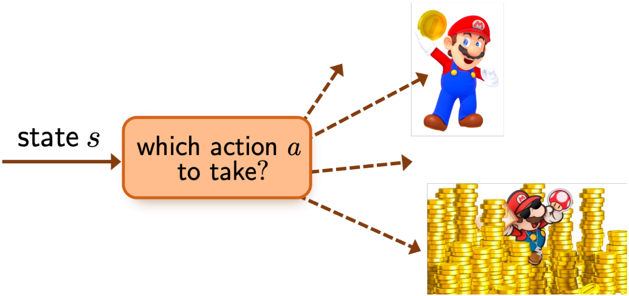
Optimal policy and optimal value



optimal policy $\pi^\star$: maximizing value function $\max_\pi V^\pi$

- optimal value / Q function: $V^\star := V^{\pi^\star}$, $Q^\star := Q^{\pi^\star}$

Optimal policy and optimal value



optimal policy $\pi^\star$: maximizing value function $\max_\pi V^\pi$

- optimal value / Q function: $V^\star := V^{\pi^\star}$, $Q^\star := Q^{\pi^\star}$
- How to find this $\pi^\star$?

Basic dynamic programming algorithms when MDP specification is known

Policy evaluation: Given MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, P, \gamma)$ and policy $\pi : \mathcal{S} \mapsto \mathcal{A}$, how good is π? (i.e., how to compute $V^\pi(s), \ \forall s$?)

Policy evaluation: Given MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, P, \gamma)$ and policy $\pi : \mathcal{S} \mapsto \mathcal{A}$, how good is π? (i.e., how to compute $V^{\pi}(s), \ \forall s$?)

Possible scheme:

- execute policy evaluation for each π
- find the optimal one

Policy evaluation: Bellman's consistency equation

- V^{π} / Q^{π}: value / action-value function under policy π

Policy evaluation: Bellman's consistency equation

- V^π / Q^π: value / action-value function under policy π

Bellman's consistency equation

$$V^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q^\pi(s,a)\right]$$

$$Q^\pi(s,a) = \underbrace{r(s,a)}_{\text{immediate reward}} + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\Big[\underbrace{V^\pi(s')}_{\text{next state's value}}\Big]$$

Richard Bellman

Policy evaluation: Bellman's consistency equation

- V^π / Q^π: value / action-value function under policy π

Bellman's consistency equation

$$V^\pi(s) = \mathbb{E}_{a \sim \pi(\cdot|s)}\left[Q^\pi(s,a)\right]$$

$$Q^\pi(s,a) = \underbrace{r(s,a)}_{\text{immediate reward}} + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)}\Big[\underbrace{V^\pi(s')}_{\text{next state's value}}\Big]$$

- one-step look-ahead

Richard Bellman

Policy evaluation: Bellman's consistency equation

- V^π / Q^π: value / action-value function under policy π

Bellman's consistency equation

$$V^\pi(s) = \mathbb{E}_{a \sim \pi(\cdot|s)}\left[Q^\pi(s,a)\right]$$

$$Q^\pi(s,a) = \underbrace{r(s,a)}_{\text{immediate reward}} + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)}\Big[\underbrace{V^\pi(s')}_{\text{next state's value}}\Big]$$

- one-step look-ahead
- let P^π be the state-action transition matrix induced by π:

$$Q^\pi = r + \gamma P^\pi Q^\pi \quad \Longrightarrow \quad Q^\pi = (I - \gamma P^\pi)^{-1} r$$

Richard Bellman

Optimal policy $\pi^\star$: Bellman's optimality principle

Bellman operator

$$\mathcal{T}(Q)(s,a) := \underbrace{r(s,a)}_{\text{immediate reward}} + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \underbrace{\left[\max_{a' \in \mathcal{A}} Q(s',a')\right]}_{\text{next state's value}}$$

- one-step look-ahead

Optimal policy $\pi^\star$: Bellman's optimality principle

Bellman operator

$$\mathcal{T}(Q)(s,a) := \underbrace{r(s,a)}_{\text{immediate reward}} + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \underbrace{\left[\max_{a' \in \mathcal{A}} Q(s',a')\right]}_{\text{next state's value}}$$

- one-step look-ahead

Bellman equation: $Q^\star$ is *unique* solution to

$$\mathcal{T}(Q^\star) = Q^\star$$

γ**-contraction of Bellman operator:**

$$\|\mathcal{T}(Q_1) - \mathcal{T}(Q_2)\|_\infty \le \gamma \|Q_1 - Q_2\|_\infty$$

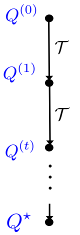
Richard Bellman

Two dynamic programming algorithms

Value iteration (VI)

For $t = 0, 1, \ldots,$

$$Q^{(t+1)} = \mathcal{T}(Q^{(t)})$$

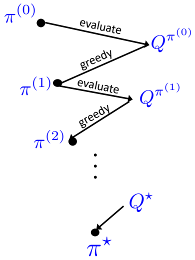


Policy iteration (PI)

For $t = 0, 1, \ldots,$

policy evaluation: $Q^{(t)} = Q^{\pi^{(t)}}$

policy improvement: $\pi^{(t+1)}(s) = \arg\max_{a \in \mathcal{A}} Q^{(t)}(s, a)$



Iteration complexity

Theorem (Linear convergence of policy/value iteration)

$$\|Q^{(t)} - Q^\star\|_\infty \le \gamma^t \|Q^{(0)} - Q^\star\|_\infty$$

Iteration complexity

Theorem (Linear convergence of policy/value iteration)

$$\big\|Q^{(t)} - Q^\star\big\|_\infty \leq \gamma^t \big\|Q^{(0)} - Q^\star\big\|_\infty$$

Implications: to achieve $\|Q^{(t)} - Q^\star\|_\infty \leq \varepsilon$, it takes no more than

$$\frac{1}{1-\gamma} \log\left(\frac{\|Q^{(0)} - Q^\star\|_\infty}{\varepsilon}\right) \quad \text{iterations}$$

Iteration complexity

Theorem (Linear convergence of policy/value iteration)

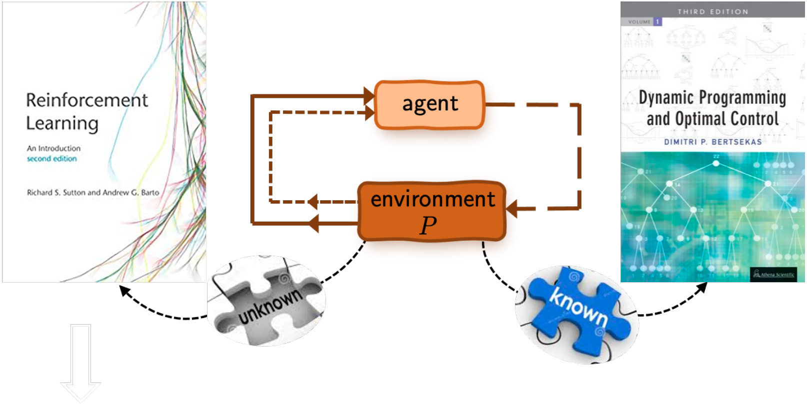
$$\|Q^{(t)} - Q^{\star}\|_{\infty} \leq \gamma^{t} \|Q^{(0)} - Q^{\star}\|_{\infty}$$

Implications: to achieve $\|Q^{(t)} - Q^{\star}\|_{\infty} \leq \varepsilon$, it takes no more than

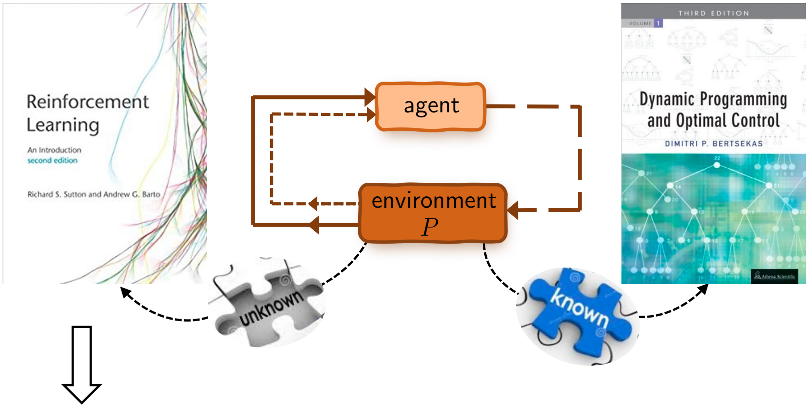
$$\frac{1}{1-\gamma} \log \left(\frac{\|Q^{(0)} - Q^{\star}\|_{\infty}}{\varepsilon} \right) \quad \text{iterations}$$

Linear convergence at a **dimension-free** rate!

When the model is unknown . . .

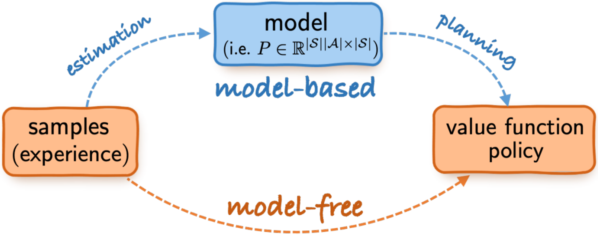


When the model is unknown . . .



Need to learn optimal policy from samples w/o model specification

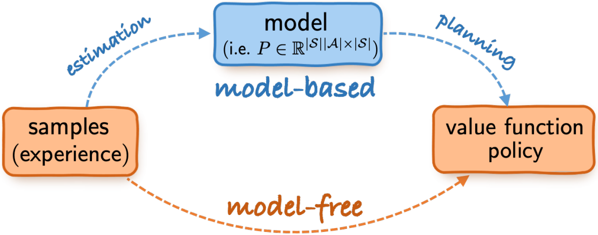
Three approaches



Model-based approach ("plug-in")

1. build an empirical estimate $\widehat{P}$ for P
2. planning based on the empirical $\widehat{P}$

Three approaches



Model-based approach ("plug-in")

1. build an empirical estimate $\widehat{P}$ for P
2. planning based on the empirical $\widehat{P}$

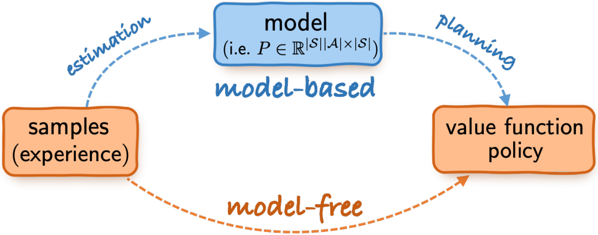
Value-based approach

— learning w/o estimating the model explicitly

Policy-based approach

— optimization in the space of policies

Three approaches



Model-based approach ("plug-in")

1. build an empirical estimate $\widehat{P}$ for P
2. planning based on the empirical $\widehat{P}$

Value-based approach

— learning w/o estimating the model explicitly

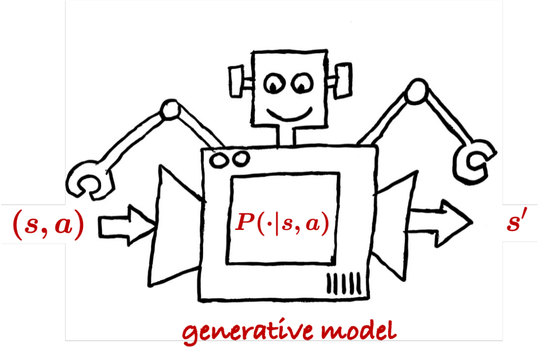
Policy-based approach

— optimization in the space of policies

Model-based RL (a "plug-in" approach)

A generative model / simulator

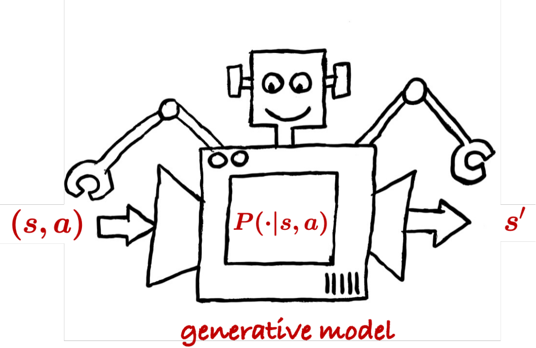
— *Kearns and Singh, 1999*



- **sampling:** for each (s, a), collect N samples $\{(s, a, s'_{(i)})\}_{1 \le i \le N}$

A generative model / simulator

— *Kearns and Singh, 1999*



- **sampling:** for each (s, a), collect N samples $\{(s, a, s'_{(i)})\}_{1 \le i \le N}$
- construct $\widehat{\pi}$ based on samples (in total $|\mathcal{S}||\mathcal{A}| \times N$)

ℓ_∞**-sample complexity:** how many samples are required to learn an $\underbrace{\varepsilon\text{-optimal policy}}_{\forall s:\ V^{\hat{\pi}}(s) \geq V^\star(s) - \varepsilon}$?

An incomplete list of works

- Kearns and Singh, 1999
- Kakade, 2003
- Kearns et al., 2002
- Azar et al., 2012
- Azar et al., 2013
- Sidford et al., 2018a, 2018b
- Wang, 2019
- Agarwal et al., 2019
- Wainwright, 2019a, 2019b
- Pananjady and Wainwright, 2019
- Yang and Wang, 2019
- Khamaru et al., 2020
- Mou et al., 2020
- Cui and Yang, 2021
- ...

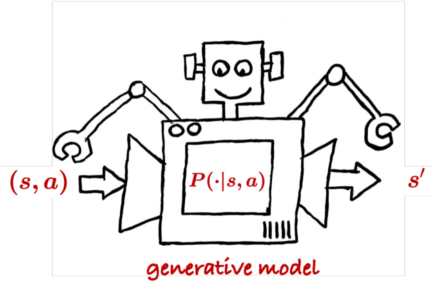
An even shorter list of prior art

algorithm	sample size range	sample complexity	ε-range
Empirical QVI Azar et al., 2013	$\big[\frac{\|\mathcal{S}\|^2\|\mathcal{A}\|}{(1-\gamma)^2}, \infty\big)$	$\frac{\|\mathcal{S}\|\|\mathcal{A}\|}{(1-\gamma)^3\varepsilon^2}$	$(0, \frac{1}{\sqrt{(1-\gamma)\|\mathcal{S}\|}}]$
Sublinear randomized VI Sidford et al., 2018b	$\big[\frac{\|\mathcal{S}\|\|\mathcal{A}\|}{(1-\gamma)^2}, \infty\big)$	$\frac{\|\mathcal{S}\|\|\mathcal{A}\|}{(1-\gamma)^4\varepsilon^2}$	$\big(0, \frac{1}{1-\gamma}\big]$
Variance-reduced QVI Sidford et al., 2018a	$\big[\frac{\|\mathcal{S}\|\|\mathcal{A}\|}{(1-\gamma)^3}, \infty\big)$	$\frac{\|\mathcal{S}\|\|\mathcal{A}\|}{(1-\gamma)^3\varepsilon^2}$	$(0, 1]$
Randomized primal-dual Wang 2019	$\big[\frac{\|\mathcal{S}\|\|\mathcal{A}\|}{(1-\gamma)^2}, \infty\big)$	$\frac{\|\mathcal{S}\|\|\mathcal{A}\|}{(1-\gamma)^4\varepsilon^2}$	$(0, \frac{1}{1-\gamma}]$
Empirical MDP + planning Agarwal et al., 2019	$\big[\frac{\|\mathcal{S}\|\|\mathcal{A}\|}{(1-\gamma)^2}, \infty\big)$	$\frac{\|\mathcal{S}\|\|\mathcal{A}\|}{(1-\gamma)^3\varepsilon^2}$	$(0, \frac{1}{\sqrt{1-\gamma}}]$

important parameters $\Longrightarrow$

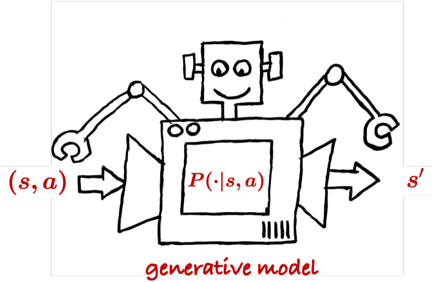
- # states $|\mathcal{S}|$, # actions $|\mathcal{A}|$
- the discounted complexity $\frac{1}{1-\gamma}$
- approximation error $\varepsilon \in (0, \frac{1}{1-\gamma}]$

Model estimation



Sampling: for each (s, a), collect N ind. samples $\{(s, a, s'_{(i)})\}_{1 \le i \le N}$

Model estimation



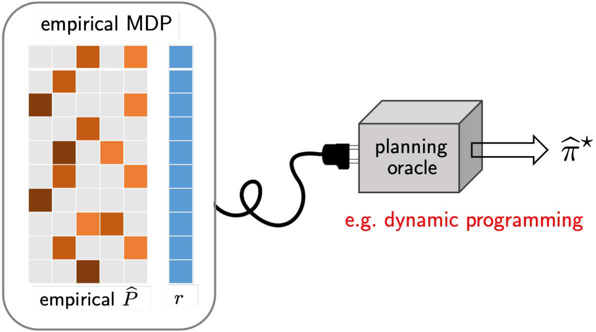
Sampling: for each (s,a), collect N ind. samples $\{(s,a,s'_{(i)})\}_{1\le i\le N}$

Empirical estimates:

$$\widehat{P}(s'|s,a) = \underbrace{\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{s'_{(i)} = s'\}}_{\text{empirical frequency}}$$

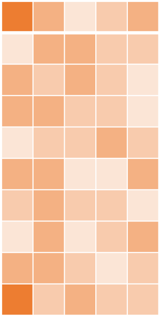
Empirical MDP + planning

— Azar et al., 2013, Agarwal et al., 2019

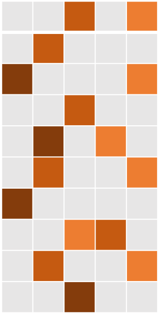


Find policy (using, e.g., policy iteration) based on the empirical MDP ($(\widehat{P}, r)$) (*empirical maximizer*)

Challenges in the sample-starved regime



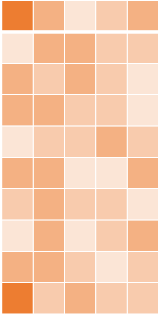
truth: $P \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|}$



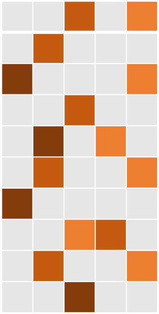
empirical estimate: $\widehat{P}$

- Can't recover P faithfully if sample size $\ll |\mathcal{S}|^2|\mathcal{A}|$!

Challenges in the sample-starved regime



truth: $P \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|}$



empirical estimate: $\widehat{P}$

- Can't recover P faithfully if sample size $\ll |\mathcal{S}|^2|\mathcal{A}|$!
- Can we trust our policy estimate when reliable model estimation is infeasible?

ℓ_∞-based sample complexity

Theorem (Agarwal, Kakade, Yang '19)

For any $0 < \varepsilon \leq \frac{1}{\sqrt{1-\gamma}}$, *the optimal policy* $\widehat{\pi}^\star$ *of empirical MDP achieves*

$$\|V^{\widehat{\pi}^\star} - V^\star\|_\infty \leq \varepsilon$$

with high prob., with sample complexity at most

$$\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\right)$$

ℓ_∞-based sample complexity

Theorem (Agarwal, Kakade, Yang '19)

For any $0 < \varepsilon \leq \frac{1}{\sqrt{1-\gamma}}$, *the optimal policy* $\widehat{\pi}^\star$ *of empirical MDP achieves*

$$\|V^{\widehat{\pi}^\star} - V^\star\|_\infty \leq \varepsilon$$

with high prob., with sample complexity at most

$$\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\right)$$

- matches minimax lower bound: $\widetilde{\Omega}(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2})$ when $\varepsilon \leq \frac{1}{\sqrt{1-\gamma}}$ (equivalently, when sample size exceeds $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}$) Azar et al., 2013

ℓ_∞-based sample complexity

Theorem (Agarwal, Kakade, Yang '19)

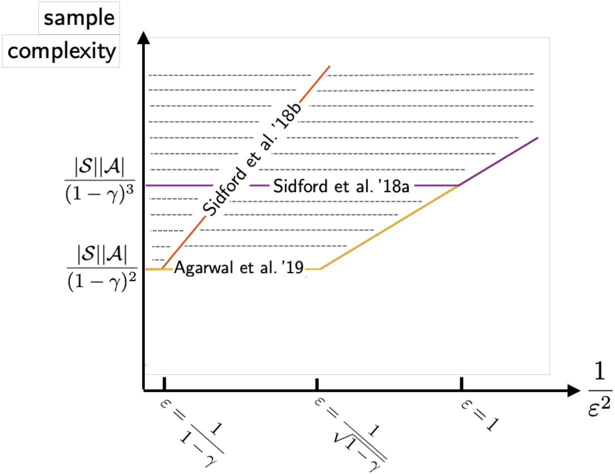
For any $0 < \varepsilon \leq \frac{1}{\sqrt{1-\gamma}}$, the optimal policy $\widehat{\pi}^\star$ of empirical MDP achieves

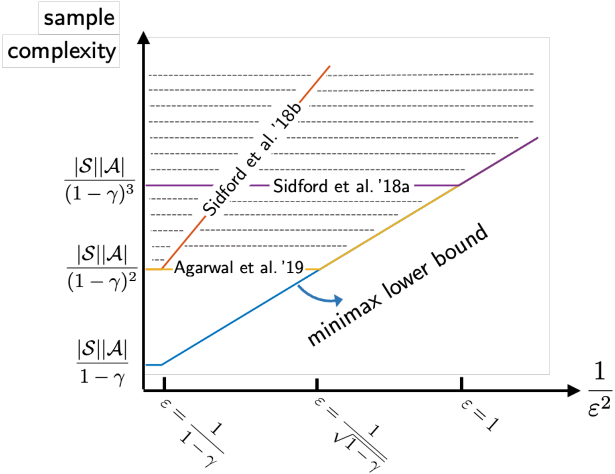
$$\|V^{\widehat{\pi}^\star} - V^\star\|_\infty \leq \varepsilon$$

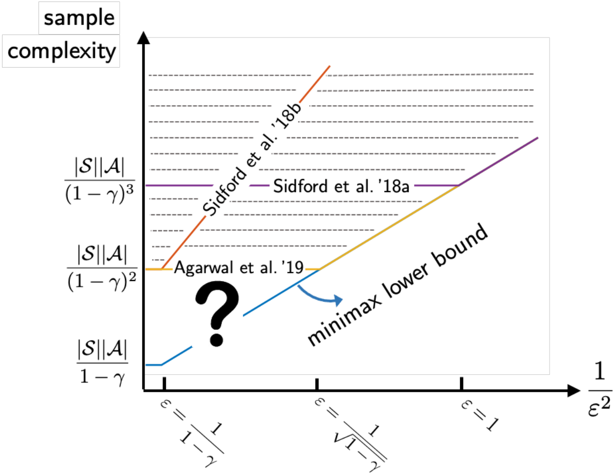
with high prob., with sample complexity at most

$$\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\right)$$

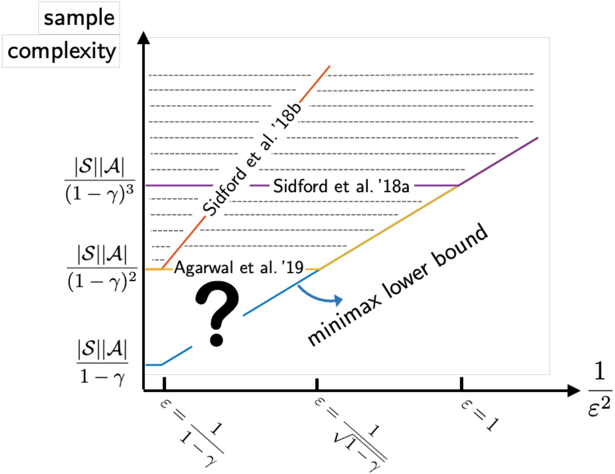
- matches minimax lower bound: $\widetilde{\Omega}(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2})$ when $\varepsilon \leq \frac{1}{\sqrt{1-\gamma}}$ (equivalently, when sample size exceeds $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}$) Azar et al., 2013
- established upon leave-one-out analysis framework


sample
complexity
$\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3}$
$\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}$
Sidford et al. '18b
Sidford et al. '18a
Agarwal et al. '19
$\frac{1}{\varepsilon^2}$
$\varepsilon = \frac{1}{1-\gamma}$
$\varepsilon = \frac{1}{\sqrt{1-\gamma}}$
$\varepsilon = 1$


sample
complexity
$\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3}$
$\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}$
$\frac{|\mathcal{S}||\mathcal{A}|}{1-\gamma}$
Sidford et al. '18b
Sidford et al. '18a
Agarwal et al. '19
minimax lower bound
$\frac{1}{\varepsilon^2}$
$\varepsilon = \frac{1}{1-\gamma}$
$\varepsilon = \frac{1}{\sqrt{1-\gamma}}$
$\varepsilon = 1$



Agarwal et al., 2019 still requires a burn-in sample size $\gtrsim \frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}$

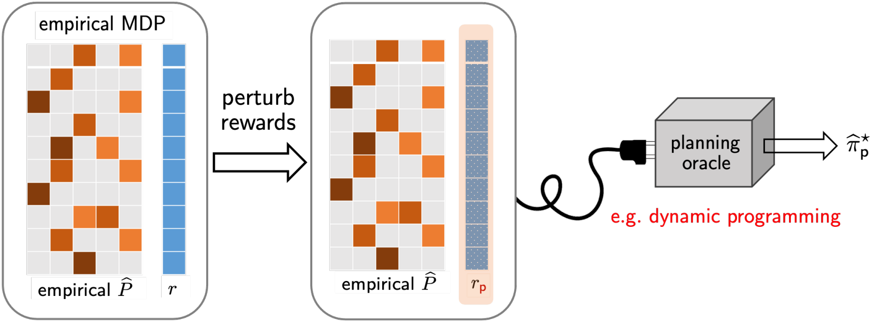


Agarwal et al., 2019 still requires a burn-in sample size $\gtrsim \frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}$

Question: is it possible to break this sample size barrier?

Perturbed model-based approach (Li et al. '20)

—*Li et al., 2020*



Find policy based on the empirical MDP with slightly perturbed rewards

Optimal ℓ_∞-based sample complexity

Theorem (Li, Wei, Chi, Chen '20)

For any $0 < \varepsilon \le \frac{1}{1-\gamma}$, *the optimal policy* $\widehat{\pi}_{\mathrm{p}}^{\star}$ *of perturbed empirical MDP achieves*

$$\|V^{\widehat{\pi}_{\mathrm{p}}^{\star}} - V^{\star}\|_\infty \le \varepsilon$$

with high prob., with sample complexity at most

$$\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\right)$$

Optimal ℓ_∞-based sample complexity

Theorem (Li, Wei, Chi, Chen '20)

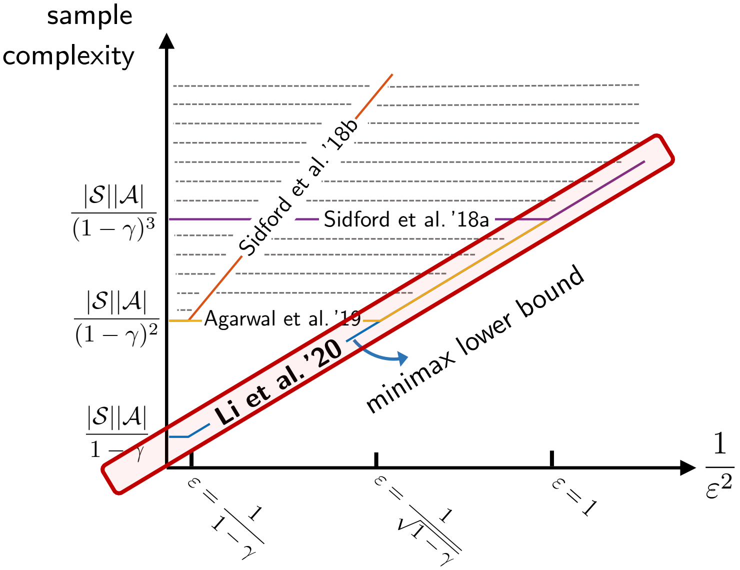
For any $0 < \varepsilon \leq \frac{1}{1-\gamma}$*, the optimal policy* $\widehat{\pi}^\star_{\mathrm{p}}$ *of perturbed empirical MDP achieves*

$$\|V^{\widehat{\pi}^\star_{\mathrm{p}}} - V^\star\|_\infty \leq \varepsilon$$

with high prob., with sample complexity at most

$$\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\right)$$

- matches minimax lower bound: $\widetilde{\Omega}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\right)$ Azar et al., 2013
- full ε-range: $\varepsilon \in \left(0, \frac{1}{1-\gamma}\right]$ $\longrightarrow$ no burn-in cost
- established upon more refined leave-one-out analysis and a perturbation argument



sample complexity
$\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3}$
$\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}$
$\frac{|\mathcal{S}||\mathcal{A}|}{1-\gamma}$
Sidford et al. '18b
Sidford et al. '18a
Agarwal et al. '19
Li et al. '20
minimax lower bound
$\frac{1}{\varepsilon^2}$
$\varepsilon = \frac{1}{1-\gamma}$
$\varepsilon = \frac{1}{\sqrt{1-\gamma}}$
$\varepsilon = 1$

A sketch of the main proof ingredients

Notation and Bellman equation

Bellman equation: $V^\pi = r_\pi + \gamma P_\pi V^\pi$

- V^π: value function under policy π
  - Bellman equation: $V^\pi = (I - \gamma P_\pi)^{-1} r_\pi$
- $\widehat{V}^\pi$: <u>empirical version</u> value function under policy π
  - Bellman equation: $\widehat{V}^\pi = (I - \gamma \widehat{P}_\pi)^{-1} r_\pi$

Notation and Bellman equation

Bellman equation: $V^\pi = r_\pi + \gamma P_\pi V^\pi$

- V^π: value function under policy π
  - Bellman equation: $V^\pi = (I - \gamma P_\pi)^{-1} r_\pi$
- $\widehat{V}^\pi$: <u>empirical version</u> value function under policy π
  - Bellman equation: $\widehat{V}^\pi = (I - \gamma \widehat{P}_\pi)^{-1} r_\pi$
- $\pi^\star$: optimal policy for V^π
- $\widehat{\pi}^\star$: optimal policy for $\widehat{V}^\pi$

Main steps

Elementary decomposition:

$$
\begin{aligned}
V^{\star} - V^{\widehat{\pi}^{\star}} &= \big(V^{\star} - \widehat{V}^{\pi^{\star}}\big) + \big(\widehat{V}^{\pi^{\star}} - \widehat{V}^{\widehat{\pi}^{\star}}\big) + \big(\widehat{V}^{\widehat{\pi}^{\star}} - V^{\widehat{\pi}^{\star}}\big) \\
&\leq \big(V^{\pi^{\star}} - \widehat{V}^{\pi^{\star}}\big) + 0 + \big(\widehat{V}^{\widehat{\pi}^{\star}} - V^{\widehat{\pi}^{\star}}\big)
\end{aligned}
$$

Main steps

Elementary decomposition:

$$\begin{aligned} V^\star - V^{\widehat{\pi}^\star} &= \big(V^\star - \widehat{V}^{\pi^\star}\big) + \big(\widehat{V}^{\pi^\star} - \widehat{V}^{\widehat{\pi}^\star}\big) + \big(\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}\big) \\ &\le \big(V^{\pi^\star} - \widehat{V}^{\pi^\star}\big) + 0 + \big(\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}\big) \end{aligned}$$

- **Step 1:** control $V^\pi - \widehat{V}^\pi$ for a fixed π (called "policy evaluation") (Bernstein inequality + a peeling argument)

Main steps

Elementary decomposition:

$$\begin{aligned} V^\star - V^{\widehat{\pi}^\star} &= \big(V^\star - \widehat{V}^{\pi^\star}\big) + \big(\widehat{V}^{\pi^\star} - \widehat{V}^{\widehat{\pi}^\star}\big) + \big(\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}\big) \\ &\le \big(V^{\pi^\star} - \widehat{V}^{\pi^\star}\big) + 0 + \big(\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}\big) \end{aligned}$$

- **Step 1:** control $V^\pi - \widehat{V}^\pi$ for a fixed π (called "policy evaluation") (Bernstein inequality + a peeling argument)
- **Step 2:** extend it to control $\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}$ ($\widehat{\pi}^\star$ depends on samples) (decouple statistical dependency)

Key idea 1: a peeling argument (for fixed policy)

First-order expansion

$$\widehat{V}^{\pi} - V^{\pi} = \gamma\big(I - \gamma P_{\pi}\big)^{-1}\big(\widehat{P}_{\pi} - P_{\pi}\big)\widehat{V}^{\pi} \qquad \text{[Agarwal et al., 2019]}$$

Key idea 1: a peeling argument (for fixed policy)

First-order expansion

$$\widehat{V}^{\pi} - V^{\pi} = \gamma\big(I - \gamma P_{\pi}\big)^{-1}\big(\widehat{P}_{\pi} - P_{\pi}\big)\widehat{V}^{\pi} \qquad \text{[Agarwal et al., 2019]}$$

Ours: higher-order expansion + Bernstein $\longrightarrow$ tighter control

$$\begin{aligned}\widehat{V}^{\pi} - V^{\pi} = \gamma\big(I - \gamma P_{\pi}\big)^{-1}\big(\widehat{P}_{\pi} - P_{\pi}\big)V^{\pi} + \\ + \gamma\big(I - \gamma P_{\pi}\big)^{-1}\big(\widehat{P}_{\pi} - P_{\pi}\big)\big(\widehat{V}^{\pi} - V^{\pi}\big)\end{aligned}$$

Bernstein's inequality: $\big|\big(\widehat{P}_{\pi} - P_{\pi}\big)V^{\pi}\big| \leq \sqrt{\frac{Var[V^{\pi}]}{N}} + \frac{\|V^{\pi}\|_{\infty}}{N}$

Key idea 1: a peeling argument (for fixed policy)

First-order expansion

$$\widehat{V}^{\pi} - V^{\pi} = \gamma\big(I - \gamma P_{\pi}\big)^{-1}\big(\widehat{P}_{\pi} - P_{\pi}\big)\widehat{V}^{\pi}$$ [Agarwal et al., 2019]

Ours: higher-order expansion + Bernstein $\longrightarrow$ tighter control

$$\begin{aligned}\widehat{V}^{\pi} - V^{\pi} = \gamma\big(I - \gamma P_{\pi}\big)^{-1}\big(\widehat{P}_{\pi} - P_{\pi}\big)V^{\pi} + \\ + \gamma^{2}\Big(\big(I - \gamma P_{\pi}\big)^{-1}\big(\widehat{P}_{\pi} - P_{\pi}\big)\Big)^{2}V^{\pi} \\ + \gamma^{3}\Big(\big(I - \gamma P_{\pi}\big)^{-1}\big(\widehat{P}_{\pi} - P_{\pi}\big)\Big)^{3}V^{\pi} \\ + \dots\end{aligned}$$

Bernstein's inequality: $|\big(\widehat{P}_{\pi} - P_{\pi}\big)V^{\pi}| \leq \sqrt{\frac{Var[V^{\pi}]}{N}} + \frac{\|V^{\pi}\|_{\infty}}{N}$

Byproduct: policy evaluation

Theorem (Li, Wei, Chi, Gu, Chen'20)

Fix any policy π. *For every* $0 < \varepsilon \leq \frac{1}{1-\gamma}$, *plug-in estimator* $\widehat{V}^{\pi}$ *obeys*

$$\|\widehat{V}^{\pi} - V^{\pi}\|_{\infty} \leq \varepsilon$$

with sample complexity at most

$$\widetilde{O}\Big(\frac{|\mathcal{S}|}{(1-\gamma)^3\varepsilon^2}\Big).$$

Byproduct: policy evaluation

Theorem (Li, Wei, Chi, Gu, Chen'20)

Fix any policy π. *For every* $0 < \varepsilon \leq \frac{1}{1-\gamma}$, *plug-in estimator* $\widehat{V}^{\pi}$ *obeys*

$$\|\widehat{V}^{\pi} - V^{\pi}\|_{\infty} \leq \varepsilon$$

with sample complexity at most

$$\widetilde{O}\Big(\frac{|\mathcal{S}|}{(1-\gamma)^3\varepsilon^2}\Big).$$

- minimax lower bound [Azar et al., 2013, Pananjady and Wainwright, 2019]

Byproduct: policy evaluation

Theorem (Li, Wei, Chi, Gu, Chen'20)

Fix any policy π. *For every* $0 < \varepsilon \le \frac{1}{1-\gamma}$, *plug-in estimator* $\widehat{V}^{\pi}$ *obeys*

$$\|\widehat{V}^{\pi} - V^{\pi}\|_{\infty} \le \varepsilon$$

with sample complexity at most

$$\widetilde{O}\Big(\frac{|\mathcal{S}|}{(1-\gamma)^3\varepsilon^2}\Big).$$

- minimax lower bound [Azar et al., 2013, Pananjady and Wainwright, 2019]
- tackle sample size barrier: prior work requires sample size $> \frac{|\mathcal{S}|}{(1-\gamma)^2}$ [Agarwal et al., 2013, Pananjady and Wainwright, 2019, Khamaru et al,, 2020]

Step 2: controlling $\widehat{V}^{\widehat{\pi}^{\star}} - V^{\widehat{\pi}^{\star}}$

A natural idea: apply our policy evaluation theory + union bound

Step 2: controlling $\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}$

A natural idea: apply our policy evaluation theory + union bound

- highly suboptimal!

Step 2: controlling $\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}$

A natural idea: apply our policy evaluation theory + union bound

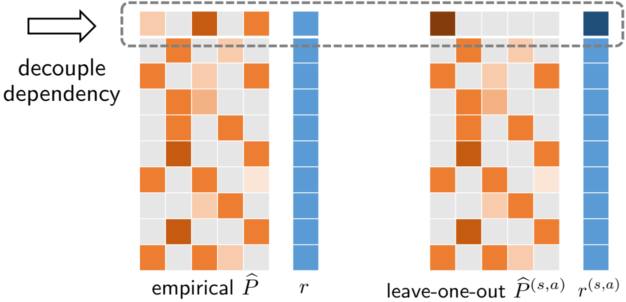
- highly suboptimal!

key idea 2: a leave-one-out argument to decouple stat. dependency btw $\widehat{\pi}$ and samples

— *inspired by [Agarwal et al., 2019] but quite different . . .*

Key idea 2: decouple dependency for $\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}$

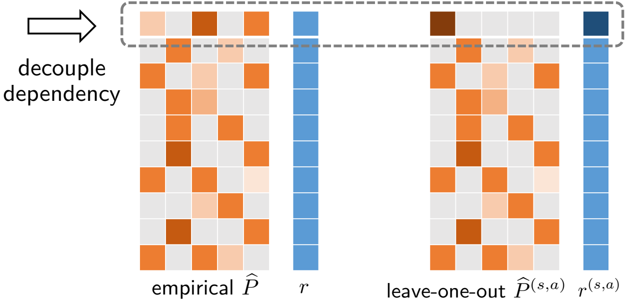
— *inspired by [Agarwal et al., 2019] but quite different . . .*



- define $\widehat{\pi}^\star_{(s,a)} \xrightarrow{\text{empirical maximizer}} \big(\widehat{P}^{(s,a)}, r^{(s,a)}\big)$

Key idea 2: decouple dependency for $\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}$

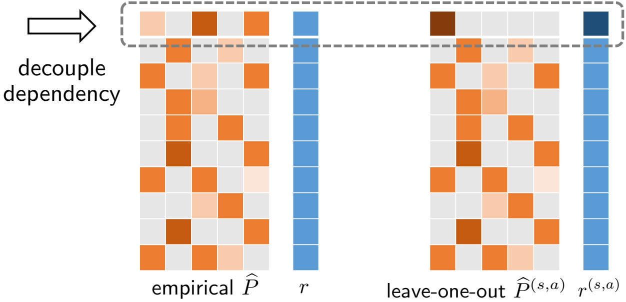
— *inspired by [Agarwal et al., 2019] but quite different . . .*



- define $\widehat{\pi}^\star_{(s,a)} \xrightarrow{\text{empirical maximizer}} \big(\widehat{P}^{(s,a)}, r^{(s,a)}\big)$
  - decouple dependency by dropping randomness in $\widehat{P}(\cdot \mid s,a)$
  - scalar $r^{(s,a)}$ ensures $\widehat{Q}^\star$ and $\widehat{V}^\star$ unchanged

Key idea 2: decouple dependency for $\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}$

— *inspired by [Agarwal et al., 2019] but quite different . . .*



- define $\widehat{\pi}^\star_{(s,a)} \xrightarrow{\text{empirical maximizer}} \big(\widehat{P}^{(s,a)}, r^{(s,a)}\big)$
- $\widehat{\pi}^\star_{(s,a)} = \widehat{\pi}^\star$ can be determined under separation condition

$$\forall s \in \mathcal{S}, \quad \widehat{Q}^\star(s, \widehat{\pi}^\star(s)) - \max_{a:\, a \neq \widehat{\pi}^\star(s)} \widehat{Q}^\star(s,a) > 0$$

Key idea 3: tie-breaking via perturbation

- How to ensure the optimal policy stand out from other policies?

$$\forall s \in \mathcal{S}, \quad \widehat{Q}^{\star}(s, \widehat{\pi}^{\star}(s)) - \max_{a: a \neq \widehat{\pi}^{\star}(s)} \widehat{Q}^{\star}(s, a) \geq \omega$$

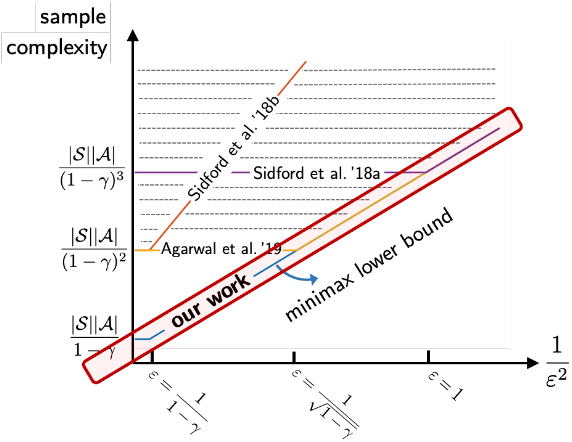
Key idea 3: tie-breaking via perturbation

- How to ensure the optimal policy stand out from other policies?

$$\forall s \in \mathcal{S}, \quad \widehat{Q}^{\star}(s, \widehat{\pi}^{\star}(s)) - \max_{a: a \neq \widehat{\pi}^{\star}(s)} \widehat{Q}^{\star}(s, a) \geq \omega$$

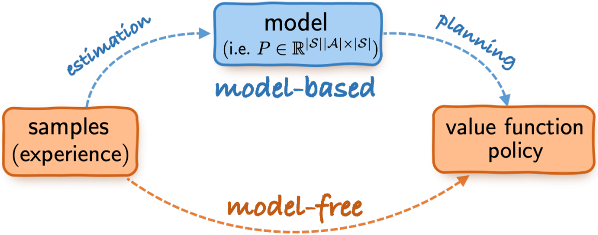
- **Solution:** slightly perturb rewards $r \implies \widehat{\pi}_{\mathrm{p}}^{\star}$
  - ensures the uniqueness of $\widehat{\pi}_{\mathrm{p}}^{\star}$
  - $V^{\widehat{\pi}_{\mathrm{p}}^{\star}} \approx V^{\widehat{\pi}^{\star}}$

Summary of model-based RL



Model-based RL is minimax optimal & does not suffer from a sample size barrier!

Three approaches



Model-based approach ("plug-in")

- build an empirical estimate $\widehat{P}$ for P
- planning based on the empirical $\widehat{P}$

Value-based approach

— learning w/o estimating the model explicitly

Policy-based approach

— optimization in the space of policies

Value-based RL (a model-free approach)

Q-learning: a stochastic approximation algorithm

Chris Watkins

Peter Dayan

$\underbrace{\text{Stochastic approximation}}_{\text{Robbins \& Monro, 1951}}$ for solving the **Bellman equation**

$$\mathcal{T}(Q) - Q = 0$$

where

$$\mathcal{T}(Q)(s,a) := \underbrace{r(s,a)}_{\text{immediate reward}} + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \Big[\underbrace{\max_{a' \in \mathcal{A}} Q(s',a')}_{\text{next state's value}} \Big].$$

Q-learning: a stochastic approximation algorithm

Chris Watkins

Peter Dayan

Stochastic approximation for solving Bellman equation $\mathcal{T}(Q) - Q = 0$

$$\underbrace{Q_{t+1}(s,a) = Q_t(s,a) + \eta_t\big(\mathcal{T}_t(Q_t)(s,a) - Q_t(s,a)\big)}_{\text{sample transition } (s,a,s')}, \quad t \geq 0$$

Q-learning: a stochastic approximation algorithm


Chris Watkins


Peter Dayan

Stochastic approximation for solving Bellman equation $\mathcal{T}(Q) - Q = 0$

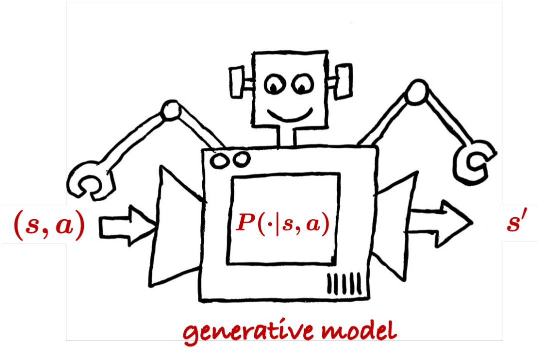
$$\underbrace{Q_{t+1}(s,a) = Q_t(s,a) + \eta_t\big(\mathcal{T}_t(Q_t)(s,a) - Q_t(s,a)\big)}_{\text{sample transition } (s,a,s')}, \quad t \geq 0$$

$$\mathcal{T}_t(Q)(s,a) = r(s,a) + \gamma \max_{a'} Q(s',a')$$

$$\mathcal{T}(Q)(s,a) = r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)}\big[\max_{a'} Q(s',a')\big]$$

A generative model / simulator

— *Kearns, Singh '99*



Each iteration, draw an independent sample (s,a,s') for given (s,a)

Synchronous Q-learning

Chris Watkins

Peter Dayan

for $t = 0, 1, \ldots, T$

  for each $(s, a) \in \mathcal{S} \times \mathcal{A}$

    draw a sample (s, a, s'), run

$$Q_{t+1}(s,a) = (1-\eta_t)Q_t(s,a) + \eta_t \Big\{ r(s,a) + \gamma \max_{a'} Q_t(s',a') \Big\}$$

synchronous: all state-action pairs are updated simultaneously

- total sample size: $T|\mathcal{S}||\mathcal{A}|$

Sample complexity of synchronous Q-learning

Theorem (Li, Cai, Chen, Wei, Chi '21)

For any $0 < \varepsilon \le 1$, synchronous Q-learning yields $\|\widehat{Q} - Q^\star\|_\infty \le \varepsilon$ with high prob. and $\mathbb{E}[\|\widehat{Q} - Q^\star\|_\infty] \le \varepsilon$, with sample size at most

$$\begin{cases} \widetilde{O}\Big(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\Big) & \text{if } |\mathcal{A}| \ge 2 \\ \widetilde{O}\Big(\frac{|\mathcal{S}|}{(1-\gamma)^3\varepsilon^2}\Big) & \text{if } |\mathcal{A}| = 1 \qquad (\textit{TD learning}) \end{cases}$$

Sample complexity of synchronous Q-learning

Theorem (Li, Cai, Chen, Wei, Chi '21)

For any $0 < \varepsilon \le 1$, synchronous Q-learning yields $\|\widehat{Q} - Q^\star\|_\infty \le \varepsilon$ with high prob. and $\mathbb{E}[\|\widehat{Q} - Q^\star\|_\infty] \le \varepsilon$, with sample size at most

$$\begin{cases} \widetilde{O}\Big(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\Big) & \text{if } |\mathcal{A}| \ge 2 \\ \widetilde{O}\Big(\frac{|\mathcal{S}|}{(1-\gamma)^3\varepsilon^2}\Big) & \text{if } |\mathcal{A}| = 1 \qquad (\textit{TD learning}) \end{cases}$$

- Covers both *constant* and *rescaled linear* learning rates:

$$\eta_t \equiv \frac{1}{1 + \frac{c_1(1-\gamma)T}{\log^2 T}} \quad \text{or} \quad \eta_t = \frac{1}{1 + \frac{c_2(1-\gamma)t}{\log^2 T}}$$

Sample complexity of synchronous Q-learning

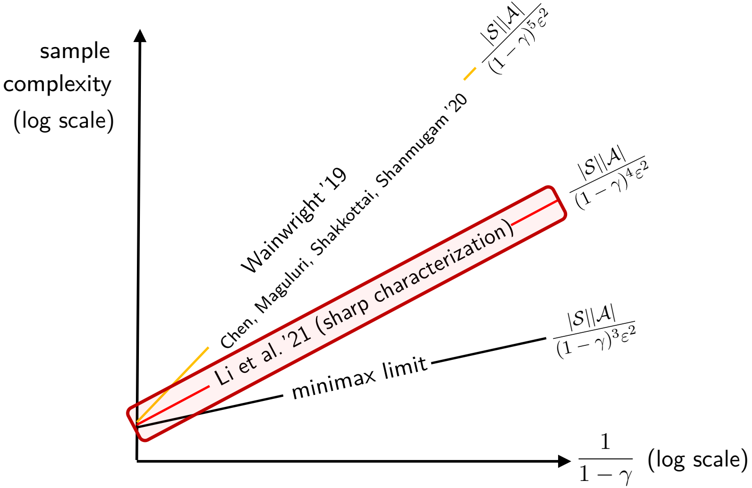
Theorem (Li, Cai, Chen, Wei, Chi '21)

For any $0 < \varepsilon \leq 1$, synchronous Q-learning yields $\|\widehat{Q} - Q^\star\|_\infty \leq \varepsilon$ with high prob. and $\mathbb{E}[\|\widehat{Q} - Q^\star\|_\infty] \leq \varepsilon$, with sample size at most

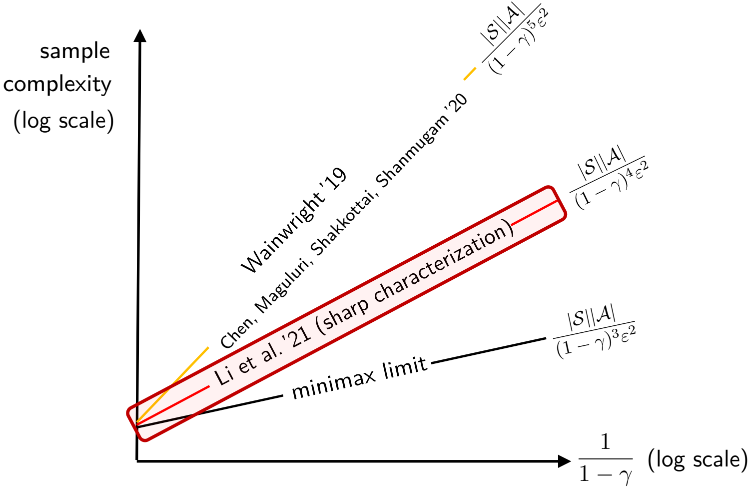
$$\begin{cases} \widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\right) & \text{if } |\mathcal{A}| \geq 2 \qquad \text{(?)} \\ \widetilde{O}\left(\frac{|\mathcal{S}|}{(1-\gamma)^3\varepsilon^2}\right) & \text{if } |\mathcal{A}| = 1 \qquad \text{(minimax optimal)} \end{cases}$$

other papers	sample complexity
Even-Dar & Mansour '03	$2^{\frac{1}{1-\gamma}} \frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\varepsilon^2}$
Beck & Srikant '12	$\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert^2}{(1-\gamma)^5\varepsilon^2}$
Wainwright '19	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^5\varepsilon^2}$
Chen, Maguluri, Shakkottai, Shanmugam '20	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^5\varepsilon^2}$

All this requires sample size at least $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}$ ($|\mathcal{A}| \geq 2$) ...



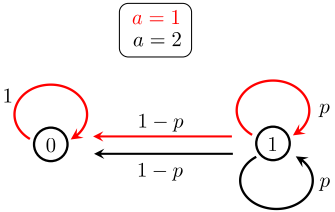
All this requires sample size at least $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}$ *($|\mathcal{A}| \geq 2$) ...*



Question: *Is Q-learning sub-optimal, or is it an analysis artifact?*

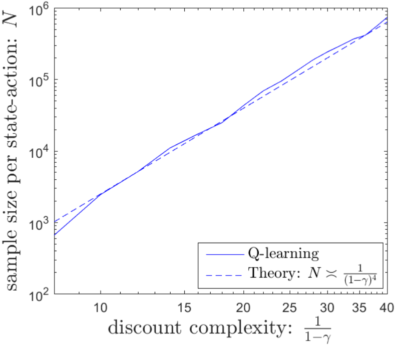
A numerical example: $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}$ samples seem necessary . . .

— *observed in Wainwright '19*



$$p = \frac{4\gamma - 1}{3\gamma}$$

$$r(0,1) = 0, \quad r(1,1) = r(1,2) = 1$$



Q-learning is NOT minimax optimal

Theorem (Li, Cai, Chen, Wei, Chi, 2021)

For any $0 < \varepsilon \leq 1$, there exists an MDP with $|\mathcal{A}| \geq 2$ such that to achieve $\|\widehat{Q} - Q^{\star}\|_{\infty} \leq \varepsilon$, synchronous Q-learning needs at least

$$\widetilde{\Omega}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\right) \quad \textit{samples}$$

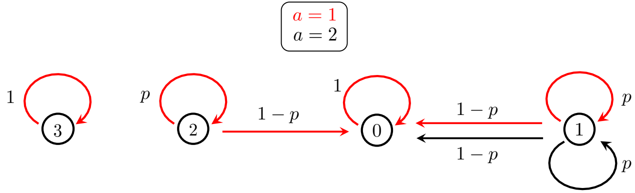
Q-learning is NOT minimax optimal

Theorem (Li, Cai, Chen, Wei, Chi, 2021)

For any $0 < \varepsilon \leq 1$, there exists an MDP with $|\mathcal{A}| \geq 2$ such that to achieve $\|\widehat{Q} - Q^\star\|_\infty \leq \varepsilon$, synchronous Q-learning needs at least

$$\widetilde{\Omega}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\right) \quad \textit{samples}$$

- Tight **algorithm-dependent** lower bound
- Holds for both constant and rescaled linear learning rates

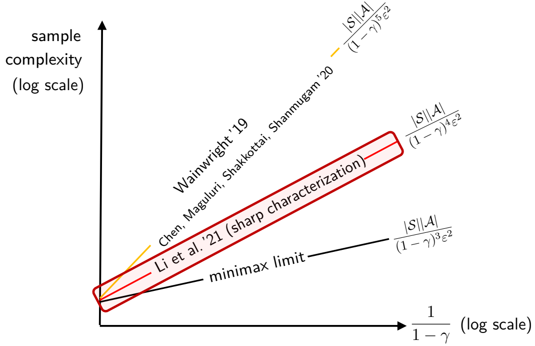


Q-learning is NOT minimax optimal

Theorem (Li, Cai, Chen, Wei, Chi, 2021)

For any $0 < \varepsilon \le 1$, there exists an MDP with $|\mathcal{A}| \ge 2$ such that to achieve $\|\widehat{Q} - Q^\star\|_\infty \le \varepsilon$, synchronous Q-learning needs at least

$$\widetilde{\Omega}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\right) \quad \textit{samples}$$



Improving sample complexity via **variance reduction**

— *a powerful idea from finite-sum stochastic optimization*

Variance-reduced Q-learning updates (Wainwright '19)

— inspired by SVRG (Johnson & Zhang '13)

$$Q_t(s,a) = (1-\eta)Q_{t-1}(s,a) + \eta\Big(\mathcal{T}_t(Q_{t-1}) \underbrace{-\mathcal{T}_t(\overline{Q}) + \widetilde{\mathcal{T}}(\overline{Q})}_{\text{use } \overline{Q} \text{ to help reduce variability}}\Big)(s,a)$$

Variance-reduced Q-learning updates (Wainwright '19)

— *inspired by SVRG (Johnson & Zhang '13)*

$$Q_t(s,a) = (1-\eta)Q_{t-1}(s,a) + \eta\Big(\mathcal{T}_t(Q_{t-1}) \underbrace{-\mathcal{T}_t(\overline{Q}) + \widetilde{\mathcal{T}}(\overline{Q})}_{\text{use } \overline{Q} \text{ to help reduce variability}}\Big)(s,a)$$

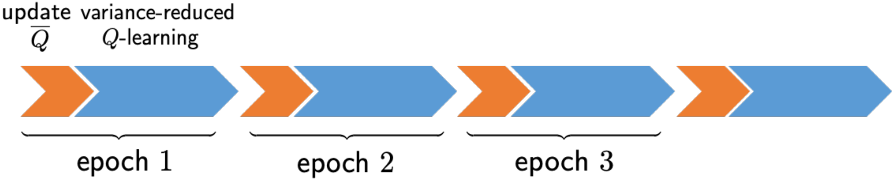
- $\overline{Q}$: some reference Q-estimate
- $\widetilde{\mathcal{T}}$: empirical Bellman operator (using a *batch* of samples)

$$\mathcal{T}_t(Q)(s,a) = r(s,a) + \gamma \max_{a'} Q(s',a')$$

$$\widetilde{\mathcal{T}}(Q)(s,a) = r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim \widetilde{P}(\cdot|s,a)} \big[\max_{a'} Q(s',a')\big]$$

An epoch-based stochastic algorithm

— *inspired by Johnson & Zhang '13*



for each epoch

1. update $\overline{Q}$ and $\widetilde{\mathcal{T}}(\overline{Q})$ (which stay fixed in the rest of the epoch)
2. run variance-reduced Q-learning updates iteratively

Sample complexity of variance-reduced Q-learning

Theorem (Wainwright '19)

For any $0 < \varepsilon \leq 1$, *sample complexity for* **variance-reduced synchronous Q-learning** *to yield* $\|\widehat{Q} - Q^{\star}\|_{\infty} \leq \varepsilon$ *is at most*

$$\widetilde{O}\Big(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\Big)$$

- allows for more aggressive learning rates

Sample complexity of variance-reduced Q-learning

Theorem (Wainwright '19)

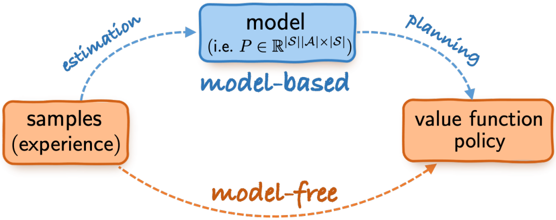
For any $0 < \varepsilon \leq 1$*, sample complexity for* **variance-reduced synchronous Q-learning** *to yield* $\|\widehat{Q} - Q^{\star}\|_{\infty} \leq \varepsilon$ *is at most*

$$\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\right)$$

- allows for more aggressive learning rates
- minimax-optimal for $0 < \varepsilon \leq 1$
  - remains suboptimal if $1 < \varepsilon < \frac{1}{1-\gamma}$

Summary of this part

- basics of MDP and DP algorithms
- break the sample size barrier using model-based approach
- obtain tight sample complexity for Q-learning



Outline (Part 2)

Four variants of our basics settings to illustrate the approaches so far:

- Offline / batch RL
- RL with Markovian samples
- Robust RL
- Multi-agent RL

Outline (Part 2)

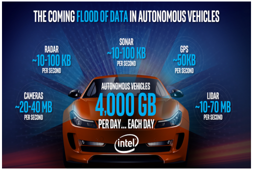
Four variants of our basics settings to illustrate the approaches so far:

- Offline / batch RL
- RL with Markovian samples
- Robust RL
- Multi-agent RL

Offline RL / batch RL

- Collecting new data might be expensive or time-consuming
- But we have already stored tons of historical data

medical records



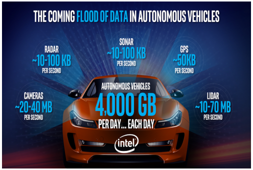
data of self-driving

clicking times of ads

Offline RL / batch RL

- Collecting new data might be expensive or time-consuming
- But we have already stored tons of historical data

medical records

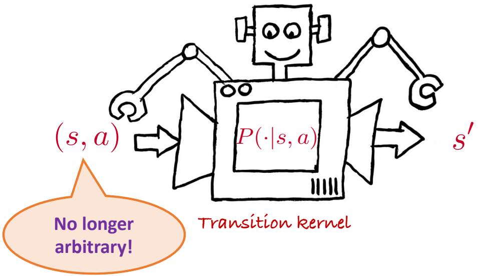


data of self-driving

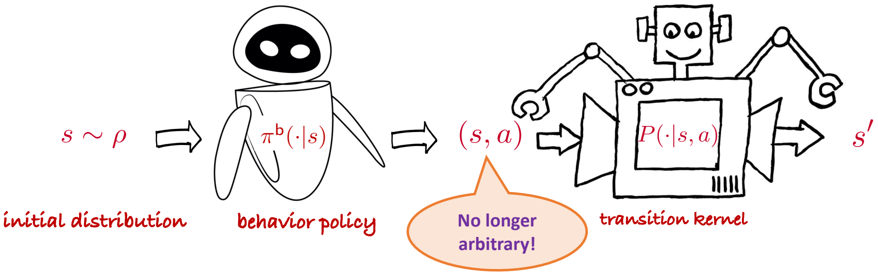
clicking times of ads

Question: Can we design algorithms based solely on historical data?

Offline RL / batch RL



Offline RL / batch RL



Offline RL / batch RL

A historical dataset $\mathcal{D} = \{(s^{(i)}, a^{(i)}, s'^{(i)})\}$**:** N independent copies of

$$s \sim \rho^{\mathsf{b}}, \qquad a \sim \pi^{\mathsf{b}}(\cdot \,|\, s), \qquad s' \sim P(\cdot \,|\, s, a)$$

for some state distribution ρ^{b} and behavior policy π^{b}

Offline RL / batch RL

A historical dataset $\mathcal{D} = \left\{\left(s^{(i)}, a^{(i)}, s'^{(i)}\right)\right\}$**:** N independent copies of

$$s \sim \rho^{\mathsf{b}}, \qquad a \sim \pi^{\mathsf{b}}(\cdot \,|\, s), \qquad s' \sim P(\cdot \,|\, s, a)$$

for some state distribution ρ^{b} and behavior policy π^{b}

Goal: given some test distribution ρ and accuracy level ε, find an ε-optimal policy $\widehat{\pi}$ based on $\mathcal{D}$ obeying

$$V^{\star}(\rho) - V^{\widehat{\pi}}(\rho) = \mathop{\mathbb{E}}_{s\sim\rho}\left[V^{\star}(s)\right] - \mathop{\mathbb{E}}_{s\sim\rho}\left[V^{\widehat{\pi}}(s)\right] \le \varepsilon$$

— *in a sample-efficient manner*

Challenges of offline RL

- **Distribution shift**:

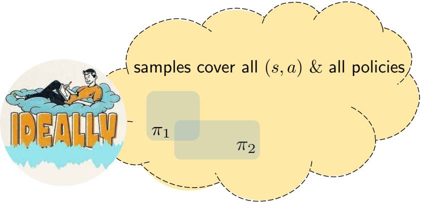
$$\text{distribution}(\mathcal{D}) \;\neq\; \text{target distribution under } \pi^\star$$

Challenges of offline RL

- **Distribution shift**:

$$\text{distribution}(\mathcal{D}) \;\neq\; \text{target distribution under } \pi^\star$$

- **Partial coverage of state-action space**:



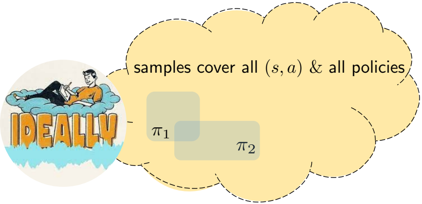
uniform coverage over entire space
(sufficiently explored)

Challenges of offline RL

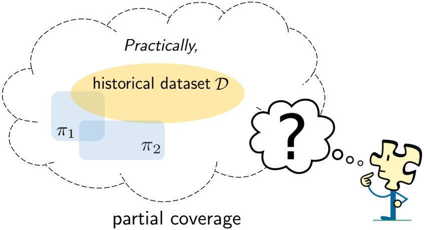
- **Distribution shift**:

$$\text{distribution}(\mathcal{D}) \neq \text{target distribution under } \pi^\star$$

- **Partial coverage of state-action space**:



uniform coverage over entire space
(sufficiently explored)



partial coverage
(inadequately explored)

How to quantify the distribution shift?

Single-policy concentrability coefficient (Rashidineiad et al.)

$$C^{\star} := \max_{s,a} \frac{d^{\pi^{\star}}(s,a)}{d^{\pi^{\mathsf{b}}}(s,a)} \geq 1$$

where $d^{\pi}(s,a) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}\big((s^t, a^t) = (s,a) \,|\, \pi\big)$ *is the state-action occupation density of policy* π*.*

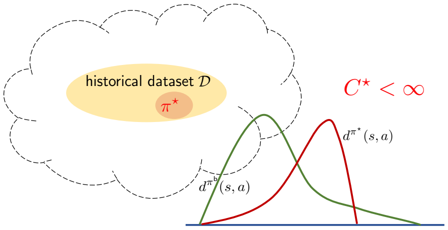
How to quantify the distribution shift?

Single-policy concentrability coefficient (Rashidineiad et al.)

$$C^\star := \max_{s,a} \frac{d^{\pi^\star}(s,a)}{d^{\pi^\mathsf{b}}(s,a)} \geq 1$$

where $d^\pi(s,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}\big((s^t,a^t)=(s,a)\,|\,\pi\big)$ *is the state-action occupation density of policy* π.

- captures distribution shift
- allows for partial coverage



How to quantify the distribution shift? — a refinement

Single-policy clipped concentrability coefficient (Li et al., '22)

$$C^{\star}_{\mathsf{clipped}} := \max_{s,a} \frac{\min\{d^{\pi^\star}(s,a), 1/S\}}{d^{\pi^{\mathsf{b}}}(s,a)} \geq 1/S$$

where $d^{\pi}(s,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}\big((s^t,a^t)=(s,a)\,|\,\pi\big)$ *is the state-action occupation density of policy* π.

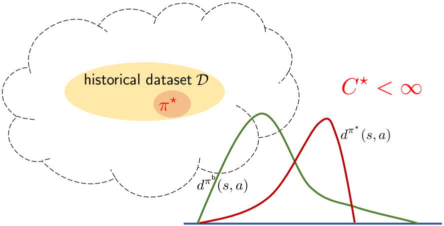
How to quantify the distribution shift? — a refinement

Single-policy clipped concentrability coefficient (Li et al., '22)

$$C^{\star}_{\mathsf{clipped}} := \max_{s,a} \frac{\min\{d^{\pi^\star}(s,a), 1/S\}}{d^{\pi^{\mathsf{b}}}(s,a)} \geq 1/S$$

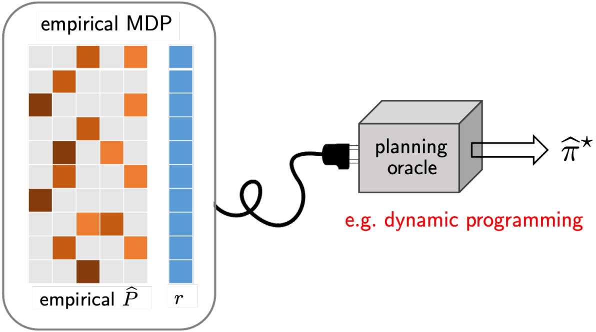
where $d^{\pi}(s,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}\big((s^t,a^t)=(s,a)\,|\,\pi\big)$ *is the state-action occupation density of policy* π.

- captures distribution shift
- allows for partial coverage
- $C^{\star}_{\mathsf{clipped}} \leq C^{\star}$



A "plug-in" model-based approach

— (Azar et al. '13, Agarwal et al. '19, Li et al. '20)



Planning (e.g., value iteration) based on the the empirical MDP $\widehat{P}$:

$$\widehat{Q}(s,a) \;\leftarrow\; r(s,a) + \gamma\big\langle \widehat{P}(\cdot \,|\, s,a), \widehat{V} \big\rangle, \quad \widehat{V}(s) = \max_a \widehat{Q}(s,a).$$

Issue: poor value estimates under partial and poor coverage.

Key idea: pessimism in the face of uncertainty

— *Jin et al. '20, Rashidinejad et al. '21, Xie et al. '21*

online

upper confidence bounds

— promote exploration of under-explored (s, a)

Key idea: pessimism in the face of uncertainty

— *Jin et al. '20, Rashidinejad et al. '21, Xie et al. '21*

online

upper confidence bounds

— promote exploration of under-explored (s, a)

offline

lower confidence bounds

— stay cautious about under-explored (s, a)

Key idea: pessimism in the face of uncertainty

— *Jin et al. '20, Rashidinejad et al. '21, Xie et al. '21*

A model-based offline algorithm: VI-LCB

1. build empirical model $\widehat{P}$
2. **(value iteration)** for $t \le \tau_{\max}$:

$$\widehat{Q}_t(s,a) \leftarrow \Big[r(s,a) + \gamma \big\langle \widehat{P}(\cdot \,|\, s,a), \widehat{V}_{t-1} \big\rangle \Big]_+$$

for all (s,a), where $\widehat{V}_t(s) = \max_a \widehat{Q}_t(s,a)$

Key idea: pessimism in the face of uncertainty

— *Jin et al. '20, Rashidinejad et al. '21, Xie et al. '21*

A model-based offline algorithm: VI-LCB

1. build empirical model $\widehat{P}$
2. **(pessimistic value iteration)** for $t \leq \tau_{\max}$:

$$\widehat{Q}_t(s,a) \leftarrow \Big[r(s,a) + \gamma \big\langle \widehat{P}(\cdot \,|\, s,a), \widehat{V}_{t-1} \big\rangle - \underbrace{b(s,a;\widehat{V}_{t-1})}_{\text{penalize poorly visited } (s,a)} \Big]_+$$

for all (s,a), where $\widehat{V}_t(s) = \max_a \widehat{Q}_t(s,a)$

Key idea: pessimism in the face of uncertainty

— *Jin et al. '20, Rashidinejad et al. '21, Xie et al. '21*

A model-based offline algorithm: VI-LCB

1. build empirical model $\widehat{P}$
2. **(pessimistic value iteration)** for $t \leq \tau_{\max}$:

$$\widehat{Q}_t(s,a) \leftarrow \Big[r(s,a) + \gamma \big\langle \widehat{P}(\cdot \,|\, s,a), \widehat{V}_{t-1} \big\rangle - \underbrace{b(s,a;\widehat{V}_{t-1})}_{\text{penalize poorly visited } (s,a)} \Big]_+$$

compared w/ prior works

- no need of variance reduction
- variance-aware penalty

Sample complexity of model-based offline RL

Theorem (Li, Shi, Chen, Chi, Wei '22)

For any $0 < \varepsilon \leq \frac{1}{1-\gamma}$, *the policy* $\widehat{\pi}$ *returned by VI-LCB achieves*

$$V^{\star}(\rho) - V^{\widehat{\pi}}(\rho) \leq \varepsilon$$

with high prob., with sample complexity at most

$$\widetilde{O}\left(\frac{SC^{\star}_{\mathsf{clipped}}}{(1-\gamma)^3\varepsilon^2}\right)$$

Sample complexity of model-based offline RL

Theorem (Li, Shi, Chen, Chi, Wei '22)

For any $0 < \varepsilon \leq \frac{1}{1-\gamma}$*, the policy* $\widehat{\pi}$ *returned by VI-LCB achieves*

$$V^{\star}(\rho) - V^{\widehat{\pi}}(\rho) \leq \varepsilon$$

with high prob., with sample complexity at most

$$\widetilde{O}\left(\frac{SC^{\star}_{\mathsf{clipped}}}{(1-\gamma)^3\varepsilon^2}\right)$$

- depends on distribution shift (as reflected by $C^{\star}_{\mathsf{clipped}}$)
- full ε-range (no burn-in cost)

Minimax optimality of model-based offline RL

Theorem (Li, Shi, Chen, Chi, Wei '22)

For any $\gamma \in [2/3, 1)$, $S \geq 2$, $C^{\star}_{\mathsf{clipped}} \geq 8\gamma/S$, *and* $0 < \varepsilon \leq \frac{1}{42(1-\gamma)}$, *there exists some MDP and batch dataset such that no algorithm succeeds if the sample size is below*

$$\widetilde{\Omega}\left(\frac{SC^{\star}_{\mathsf{clipped}}}{(1-\gamma)^3\varepsilon^2}\right).$$

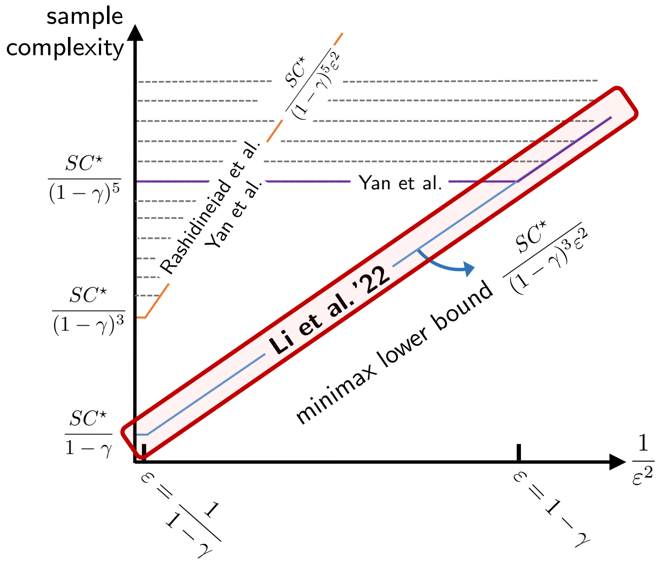
Minimax optimality of model-based offline RL

Theorem (Li, Shi, Chen, Chi, Wei '22)

For any $\gamma \in [2/3, 1)$, $S \geq 2$, $C^{\star}_{\mathsf{clipped}} \geq 8\gamma/S$, and $0 < \varepsilon \leq \frac{1}{42(1-\gamma)}$, there exists some MDP and batch dataset such that no algorithm succeeds if the sample size is below

$$\widetilde{\Omega}\left(\frac{SC^{\star}_{\mathsf{clipped}}}{(1-\gamma)^3\varepsilon^2}\right).$$

- verifies the near-minimax optimality of the pessimistic model-based algorithm
- improves upon prior results by allowing $C^{\star}_{\mathsf{clipped}} \asymp 1/S$.



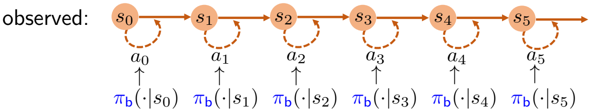
sample complexity
$\frac{SC^\star}{(1-\gamma)^5\varepsilon^2}$
Rashidinejad et al.
Yan et al.
$\frac{SC^\star}{(1-\gamma)^5}$
Yan et al.
$\frac{SC^\star}{(1-\gamma)^3}$
Li et al. '22
$\frac{SC^\star}{(1-\gamma)^3\varepsilon^2}$
minimax lower bound
$\frac{SC^\star}{1-\gamma}$
$\frac{1}{\varepsilon^2}$
$\varepsilon = \frac{1}{1-\gamma}$
$\varepsilon = 1-\gamma$

Outline (Part 2)

Four variants of our basics settings to illustrate the approaches so far:

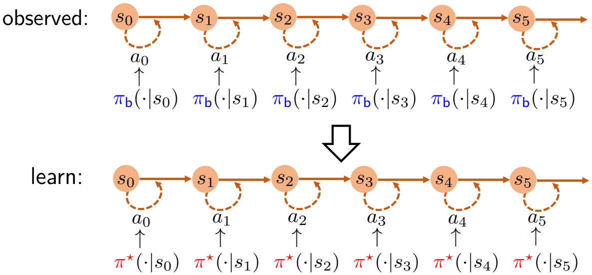
- Offline / batch RL
- RL with Markovian samples
- Robust RL
- Multi-agent RL

Markovian samples and behavior policy



Observed: $\underbrace{\{s_t, a_t, r_t\}_{t\geq 0}}_{\text{Markovian trajectory}}$ induced by behavior policy π_b

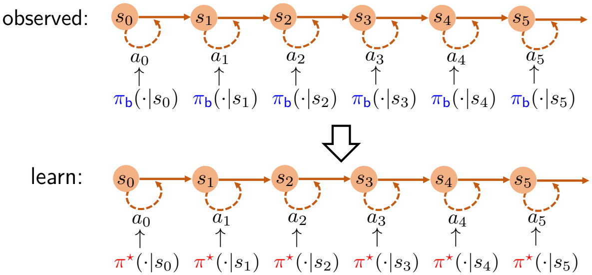
Markovian samples and behavior policy



Observed: $\underbrace{\{s_t, a_t, r_t\}_{t\geq 0}}_{\text{Markovian trajectory}}$ induced by behavior policy π_b

Goal: learn optimal value $V^\star$ and $Q^\star$ based on sample trajectory

Markovian samples and behavior policy



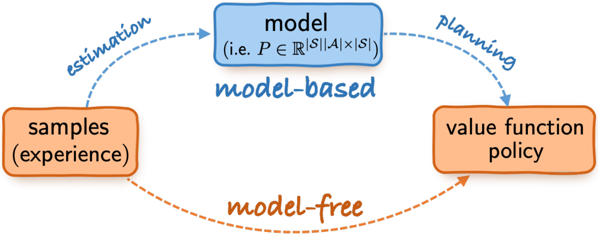
Key quantities of sample trajectory

- minimum state-action occupancy probability

$$\mu_{\min} := \min \underbrace{\mu_{\pi_b}(s,a)}_{\text{stationary distribution}}$$

- mixing time: t_{mix}

Model-based vs. model-free RL



Model-free approach (e.g. Q-learning)

— learning w/o modeling & estimating environment explicitly

Q-learning: a classical model-free algorithm

Chris Watkins

Peter Dayan

$\underbrace{\text{Stochastic approximation}}_{\text{Robbins \& Monro '51}}$ for solving **Bellman equation** $Q = \mathcal{T}(Q)$

Q-learning: a classical model-free algorithm

Chris Watkins

Peter Dayan

Stochastic approximation for solving Bellman equation $Q = \mathcal{T}(Q)$

$$\underbrace{Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \eta_t(\mathcal{T}_t(Q_t)(s_t, a_t) - Q_t(s_t, a_t))}_{\textit{only} \text{ update } (s_t, a_t)\text{-th entry}}, \quad t \geq 0$$

Q-learning: a classical model-free algorithm

Chris Watkins

Peter Dayan

Stochastic approximation for solving Bellman equation $Q = \mathcal{T}(Q)$

$$\underbrace{Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \eta_t(\mathcal{T}_t(Q_t)(s_t, a_t) - Q_t(s_t, a_t)),}_{\textit{only}\text{ update } (s_t, a_t)\text{-th entry}} \quad t \geq 0$$

$$\mathcal{T}_t(Q)(s_t, a_t) := r(s_t, a_t) + \gamma \max_{a'} Q(s_{t+1}, a')$$

$$\mathcal{T}(Q)(s, a) = r(s, a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left[\max_{a'} Q(s', a') \right]$$

Q-learning: a classical model-free algorithm

Chris Watkins

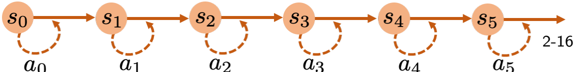
Peter Dayan

Stochastic approximation for solving Bellman equation $Q = \mathcal{T}(Q)$

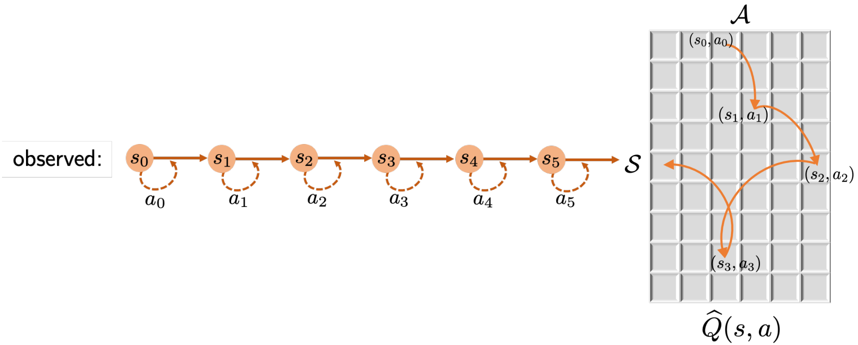
$$\underbrace{Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \eta_t(\mathcal{T}_t(Q_t)(s_t, a_t) - Q_t(s_t, a_t)),}_{\textit{only}\text{ update }(s_t, a_t)\text{-th entry}} \quad t \geq 0$$

— **asynchronous:** only a single entry is updated each iteration (resembles Markov-chain *coordinate descent*)

observed:

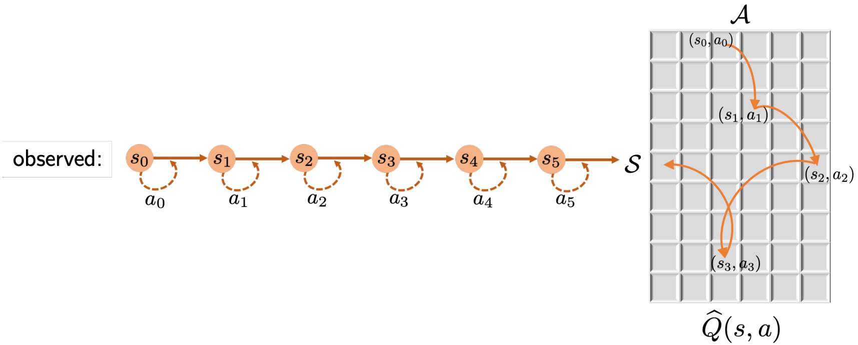


Q-learning on Markovian samples



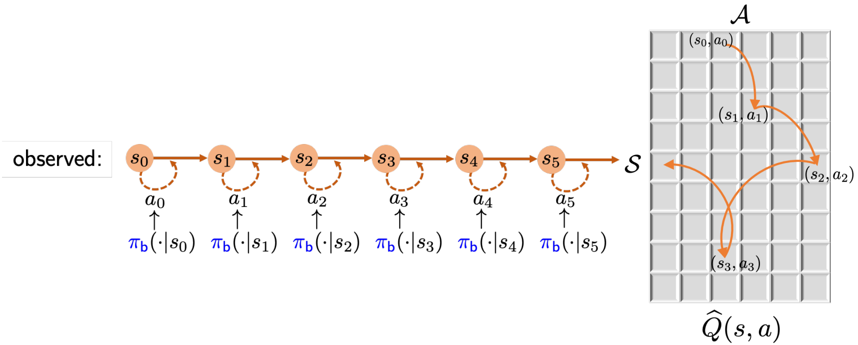
- **asynchronous:** only a single entry is updated each iteration

Q-learning on Markovian samples



- **asynchronous:** only a single entry is updated each iteration
  - resembles Markov-chain *coordinate descent*

Q-learning on Markovian samples



- **asynchronous:** only a single entry is updated each iteration
  - resembles Markov-chain *coordinate descent*
- **off-policy:** target policy $\pi^\star \neq$ behavior policy π_b

What is sample complexity of (async) Q-learning?

A highly incomplete list of works

- Watkins, Dayan '92
- Tsitsiklis '94
- Jaakkola, Jordan, Singh '94
- Szepesvári '98
- Borkar, Meyn '00
- Even-Dar, Mansour '03
- Beck, Srikant '12
- Chi, Zhu, Bubeck, Jordan '18
- Lee, He '18
- Chen, Zhang, Doan, Maguluri, Clarke '19
- Du, Lee, Mahajan, Wang '20
- Chen, Maguluri, Shakkottai, Shanmugam '20
- Qu, Wierman '20
- Devraj, Meyn '20
- Weng, Gupta, He, Ying, Srikant '20
- Li, Wei, Chi, Gu, Chen '20
- Li, Cai, Chen, Wei, Chi '21
- Chen, Maguluri, Shakkottai, Shanmugam '21
- ...

Prior art: async Q-learning

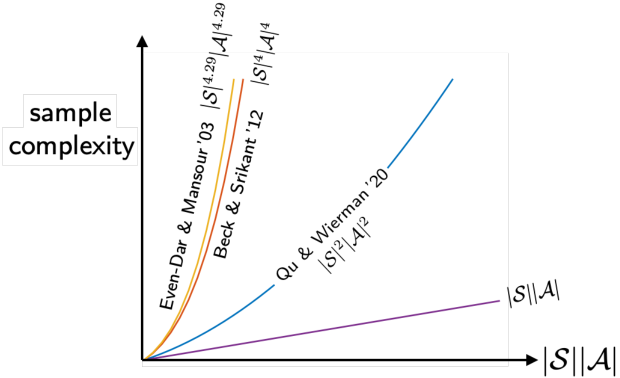
Question: how many samples are needed to ensure $\|\widehat{Q} - Q^\star\|_\infty \le \varepsilon$?

other papers	sample complexity
Even-Dar, Mansour '03	$\frac{(t_{\text{cover}})^{\frac{1}{1-\gamma}}}{(1-\gamma)^4\varepsilon^2}$
Even-Dar, Mansour '03	$\big(\frac{t_{\text{cover}}^{1+3\omega}}{(1-\gamma)^4\varepsilon^2}\big)^{\frac{1}{\omega}} + \big(\frac{t_{\text{cover}}}{1-\gamma}\big)^{\frac{1}{1-\omega}}$, $\omega \in (\frac{1}{2}, 1)$
Beck & Srikant '12	$\frac{t_{\text{cover}}^3 \lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^5\varepsilon^2}$
Qu & Wierman '20	$\frac{t_{\text{mix}}}{\mu_{\text{min}}^2(1-\gamma)^5\varepsilon^2}$
Li, Wei, Chi, Gu, Chen '20	$\frac{1}{\mu_{\text{min}}(1-\gamma)^5\varepsilon^2} + \frac{t_{\text{mix}}}{\mu_{\text{min}}(1-\gamma)}$
Chen, Maguluri, Shakkottai, Shanmugam '21	$\frac{1}{\mu_{\text{min}}^3(1-\gamma)^5\varepsilon^2} + \text{other-term}(t_{\text{mix}})$

— cover time: $t_{\text{cover}} \asymp \frac{t_{\text{mix}}}{\mu_{\text{min}}}$

Prior art: async Q-learning

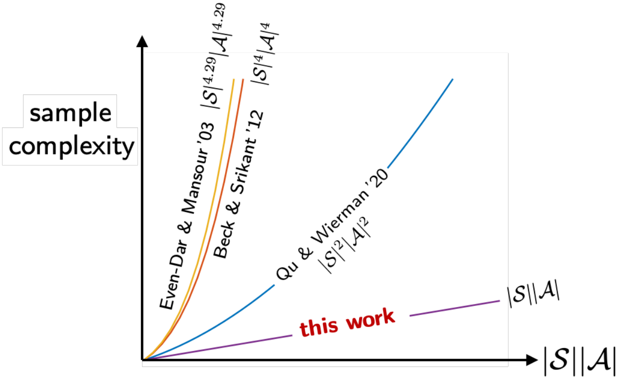
Question: how many samples are needed to ensure $\|\widehat{Q} - Q^{\star}\|_{\infty} \leq \varepsilon$?



if we take $\mu_{\mathsf{min}} \asymp \frac{1}{|\mathcal{S}||\mathcal{A}|}$, $t_{\mathsf{cover}} \asymp \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}}$

Prior art: async Q-learning

Question: how many samples are needed to ensure $\|\widehat{Q} - Q^\star\|_\infty \le \varepsilon$?



if we take $\mu_{\mathsf{min}} \asymp \frac{1}{|\mathcal{S}||\mathcal{A}|}$, $t_{\mathsf{cover}} \asymp \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}}$

All prior results require sample size of at least $t_{\mathsf{mix}}|\mathcal{S}|^2|\mathcal{A}|^2$!

Main result: ℓ_∞-based sample complexity

Theorem (Li, Wei, Chi, Gu, Chen '20)

For any $0 < \varepsilon \leq \frac{1}{1-\gamma}$*, sample complexity of async Q-learning to yield* $\|\widehat{Q} - Q^\star\|_\infty \leq \varepsilon$ *is at most (up to some log factor)*

$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^5\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}$$

Main result: ℓ_∞-based sample complexity

Theorem (Li, Wei, Chi, Gu, Chen '20)

For any $0 < \varepsilon \le \frac{1}{1-\gamma}$*, sample complexity of async Q-learning to yield* $\|\widehat{Q} - Q^\star\|_\infty \le \varepsilon$ *is at most (up to some log factor)*

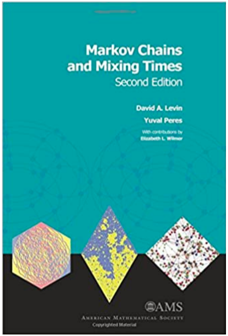
$$\frac{1}{\mu_{\min}(1-\gamma)^5\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\min}(1-\gamma)}$$

— *prior art:* $\frac{t_{\mathsf{mix}}}{\mu_{\min}^2(1-\gamma)^5\varepsilon^2}$ (Qu & Wierman'20)

- Improves upon prior art by **at least** $|\mathcal{S}||\mathcal{A}|$!

Effect of mixing time on sample complexity

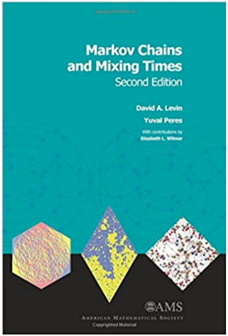
$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^5\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}$$



- reflects cost taken to reach steady state
- one-time expense (almost independent of ε)
  — it becomes amortized as algorithm runs

Effect of mixing time on sample complexity

$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^5\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}$$



- reflects cost taken to reach steady state
- one-time expense (almost independent of ε)

  — it becomes amortized as algorithm runs

— *prior art:* $\frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}^2(1-\gamma)^5\varepsilon^2}$ [Qu & Wierman '20]

Dependence on effective horizon

minimax lower bound
(Azar et al. '13)

$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^3\varepsilon^2}$$

asyn Q-learning
(ignoring dependency on t_{mix})

$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^5\varepsilon^2}$$

Dependence on effective horizon

minimax lower bound
(Azar et al. '13)

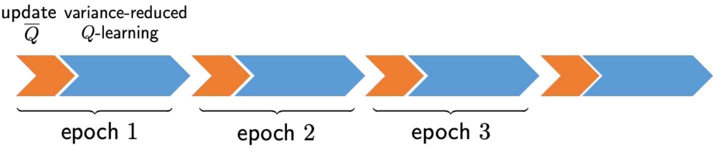
$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^3\varepsilon^2}$$

asyn Q-learning
(ignoring dependency on t_{mix})

$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^5\varepsilon^2}$$

The dependency on $\frac{1}{1-\gamma}$ can be tightened by *variance reduction*.

— *inspired by [Johnson & Zhang, 2013], [Wainwright, 2019]*



Sample complexity for variance-reduced Q-learning

Theorem (Li, Wei, Chi, Gu, Chen '20)

For any $0 < \varepsilon \leq 1$, sample complexity for **(async) variance-reduced Q-learning** *to yield $\|\widehat{Q} - Q^\star\|_\infty \leq \varepsilon$ is at most on the order of*

$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^3\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}$$

- more aggressive learning rates: $\eta_t \equiv \min\left\{\frac{\cancel{(1-\gamma)^4}(1-\gamma)^2}{\gamma^2}, \frac{1}{t_{\mathsf{mix}}}\right\}$
- minimax-optimal for $0 < \varepsilon \leq 1$

Outline (Part 2)

Four variants of our basics settings to illustrate the approaches so far:

- Offline / batch RL
- RL with Markovian samples
- Robust RL
- Multi-agent RL

Robustness and safety

(Zhou et al., 2021; Panaganti and Kalathil, 2022; Yang et al., 2022;)

Training environment $\neq$ Test environment

Sim2Real Gap: Can we learn optimal policies that are robust to model perturbations?

Uncertainty set of transition kernels: $\mathcal{U}^{\sigma}(P^{o})$

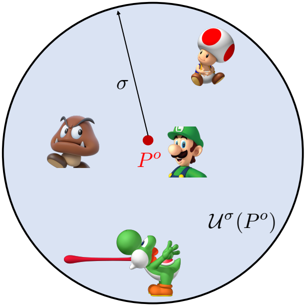
Uncertainty set with (s, a)-rectangular (Wiesemann et al. '13)

The uncertainty set is defined as a ball around the nominal transition kernel P^o ($P^o_{s,a} := P^o(\cdot \mid s, a) \in \mathbb{R}^{1\times S}$):

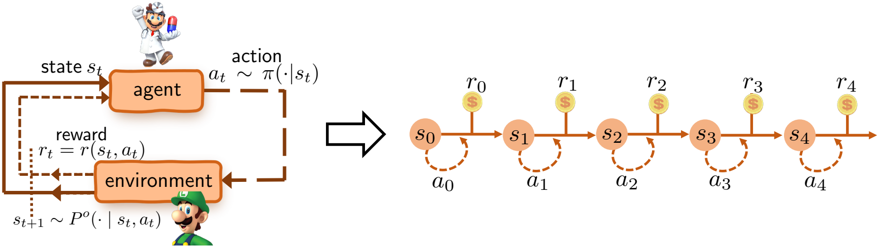
$$\mathcal{U}^{\sigma}(P^o) := \otimes \mathcal{U}^{\sigma}(P^o_{s,a}),$$

$$\mathcal{U}^{\sigma}(P^o_{s,a}) := \left\{ \mathcal{P} \in \Delta(\mathcal{S}) : \rho\left(\mathcal{P} \parallel P^o_{s,a}\right) \le \sigma \right\}.$$

- $\rho : \Delta(\mathcal{S}) \times \Delta(\mathcal{S}) \to [0, \infty]$: some distance functions (Kullback-Leibler (KL) divergence)
- $\sigma > 0$: the uncertainty level/radius
- $\otimes$: the Cartesian product



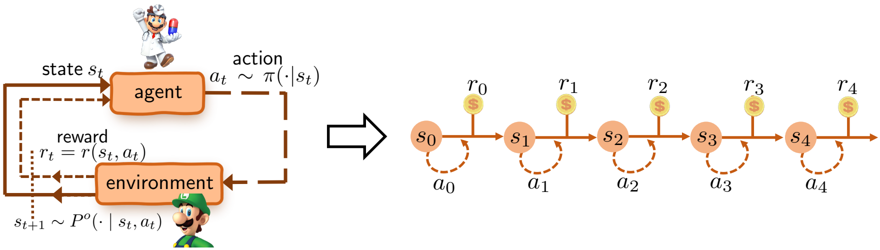
Value function: discounted infinite-horizon MDP



execute policy π to generate sample trajectory $\{(s_t, a_t)\}_{t\geq 0}$

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad V^{\pi,P}(s) := \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty} \gamma^t r\big(s_t, a_t\big) \,\Big|\, s_0 = s\right]$$

Value function: discounted infinite-horizon MDP



execute policy π to generate sample trajectory $\{(s_t, a_t)\}_{t\geq 0}$

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad V^{\pi,P}(s) := \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s\right]$$

- $\gamma \in [0,1)$: discount factor;
- P: any transition kernel

Robust value function: infinite-horizon robust MDP

- Classical value-function/Q-function:

$$V^{\pi,P}(s) := \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\,\middle|\, s_0=s\right]$$

$$Q^{\pi,P}(s,a) := \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\,\middle|\, s_0=s, a_0=a\right]$$

Robust value function: infinite-horizon robust MDP

- Classical value-function/Q-function:

$$V^{\pi,P}(s) := \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty}\gamma^t r\big(s_t,a_t\big)\,\Big|\, s_0 = s\right]$$

$$Q^{\pi,P}(s,a) := \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\,\Big|\, s_0 = s, a_0 = a\right]$$

- **Robust value function/Q-function:**

$$V^{\pi,\sigma}(s) := \inf_{P\in\mathcal{U}^{\sigma}(P^o)} V^{\pi,P}(s), \qquad Q^{\pi,\sigma}(s,a) := \inf_{P\in\mathcal{U}^{\sigma}(P^o)} Q^{\pi,P}(s,a)$$

Robust value function: infinite-horizon robust MDP

- Classical value-function/Q-function:

$$V^{\pi,P}(s) := \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty}\gamma^t r\big(s_t,a_t\big)\,\Big|\, s_0 = s\right]$$

$$Q^{\pi,P}(s,a) := \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\,\Big|\, s_0 = s, a_0 = a\right]$$

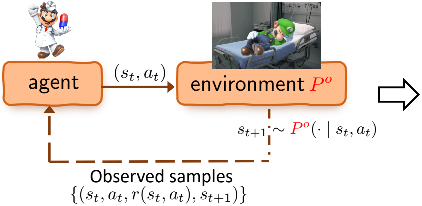
- **Robust value function/Q-function:**

$$V^{\pi,\sigma}(s) := \inf_{P\in\mathcal{U}^{\sigma}(P^{o})} V^{\pi,P}(s), \qquad Q^{\pi,\sigma}(s,a) := \inf_{P\in\mathcal{U}^{\sigma}(P^{o})} Q^{\pi,P}(s,a)$$

- **Optimal robust policy $\pi^\star$:** $\arg\max_{\pi} V^{\pi,\sigma}$
- **Optimal robust values:** $V^{\star,\sigma} := V^{\pi^\star,\sigma} = \max_{\pi} V^{\pi,\sigma}$

Classical MDP v.s robust MDP (RMDP)

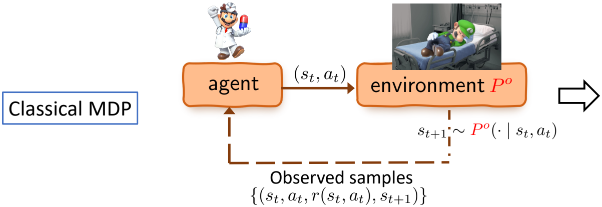
Classical MDP



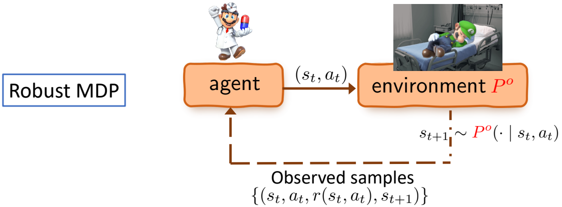
Goal: learn an optimal policy for $\mathcal{M} = \{P^o, r, \mathcal{S}, \mathcal{A}, \gamma\}$

Classical MDP v.s robust MDP (RMDP)

- Robust MDP: $\mathcal{M}_{\text{rob}} = \{\mathcal{U}(P^o), r, \mathcal{S}, \mathcal{A}, \gamma\}$
  - P^o: unknown nominal transition kernel



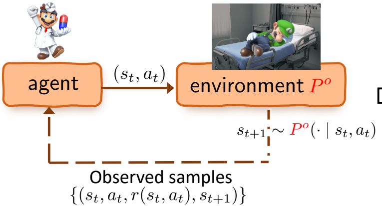
Goal: learn an optimal policy for $\mathcal{M} = \{P^o, r, \mathcal{S}, \mathcal{A}, \gamma\}$



Classical MDP v.s robust MDP (RMDP)

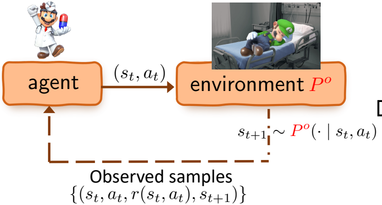
- Robust MDP: $\mathcal{M}_{\mathsf{rob}} = \{\mathcal{U}(P^o), r, \mathcal{S}, \mathcal{A}, \gamma\}$
  - P^o: unknown nominal transition kernel
  - $\mathcal{U}(P^o)$: an uncertainty set around P^o

Classical MDP

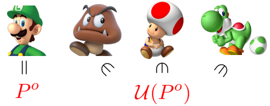


Goal: learn an optimal policy for $\mathcal{M} = \{P^o, r, \mathcal{S}, \mathcal{A}, \gamma\}$

Robust MDP



Goal: Learn an optimal policy for $\mathcal{M}_{\mathsf{rob}} = \{\mathcal{U}(P^o), r, \mathcal{S}, \mathcal{A}, \gamma\}$



Robust Bellman's optimality equation

(Iyengar. '05, Nilim and El Ghaoui. '05)

Robust Bellman's optimality equation: the optimal robust policy $\pi^\star$ and optimal robust value $V^{\star,\sigma} := V^{\pi^\star,\sigma}$ satisfy

$$
\begin{aligned}
Q^{\star,\sigma}(s,a) &= r(s,a) + \gamma \inf_{P_{s,a} \in \mathcal{U}^\sigma\left(P^o_{s,a}\right)} \left\langle P_{s,a}, V^{\star,\sigma} \right\rangle, \\
V^{\star,\sigma}(s) &= \max_a Q^{\star,\sigma}(s,a)
\end{aligned}
$$

Robust Bellman's optimality equation

(Iyengar. '05, Nilim and El Ghaoui. '05)

Robust Bellman's optimality equation: the optimal robust policy $\pi^\star$ and optimal robust value $V^{\star,\sigma} := V^{\pi^\star,\sigma}$ satisfy

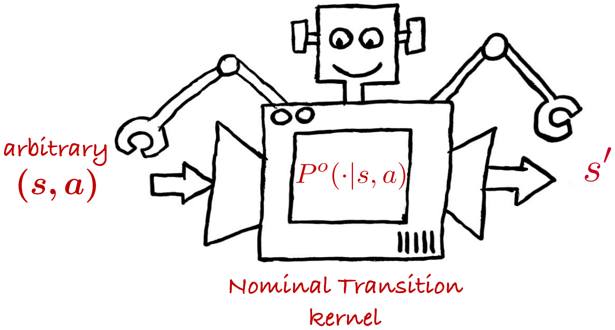
$$Q^{\star,\sigma}(s,a) = r(s,a) + \gamma \inf_{P_{s,a} \in \mathcal{U}^\sigma\left(P^o_{s,a}\right)} \left\langle P_{s,a}, V^{\star,\sigma} \right\rangle,$$
$$V^{\star,\sigma}(s) = \max_a Q^{\star,\sigma}(s,a)$$

Robust value iteration:

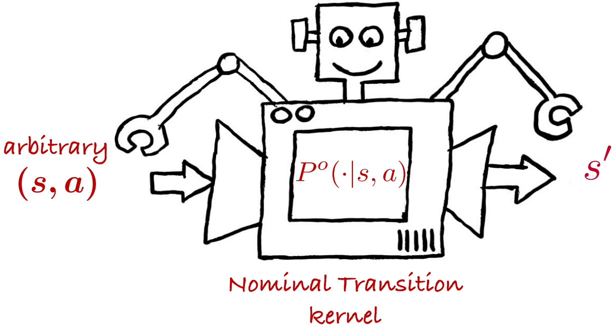
$$Q(s,a) \leftarrow r(s,a) + \gamma \inf_{P_{s,a} \in \mathcal{U}^\sigma\left(P^o_{s,a}\right)} \langle P_{s,a}, V \rangle,$$

where $V(s) = \max_a Q(s,a)$.

Learning distributionally robust MDPs



Learning distributionally robust MDPs

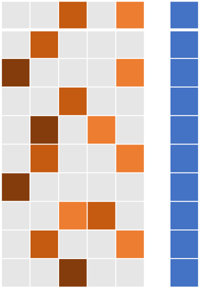


Goal of robust RL: given $\mathcal{D} := \{(s_i, a_i, s'_i)\}_{i=1}^N$ from the *nominal* environment P^0, find an ε-optimal robust policy $\widehat{\pi}$ obeying

$$V^{\star,\sigma}(\rho) - V^{\widehat{\pi},\sigma}(\rho) \leq \varepsilon$$

— *in a sample-efficient manner*

A curious question

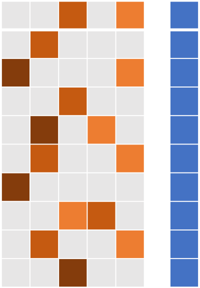


empirical MDP

Learn the optimal policy of the nominal MDP?

Learn the **robust** policy around the nominal MDP?

A curious question



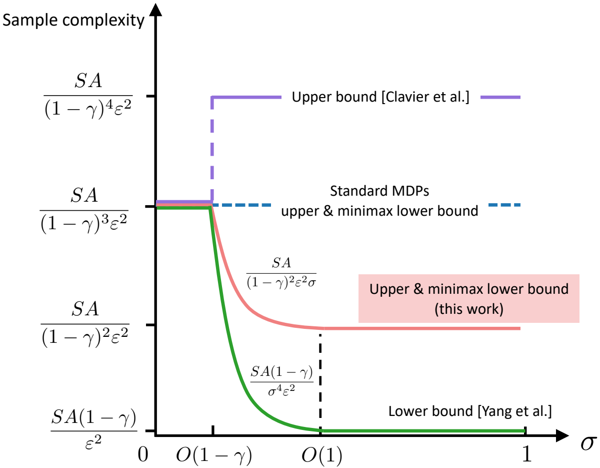
empirical MDP

Learn the optimal policy of the nominal MDP?

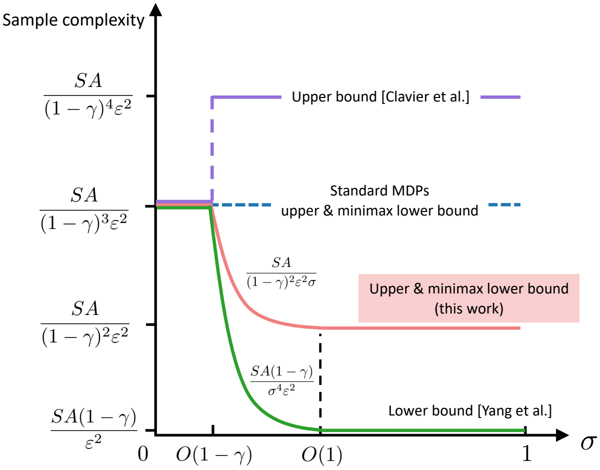
Learn the **robust** policy around the nominal MDP?

Robustness-statistical trade-off? Is there a statistical premium that one needs to pay in quest of additional robustness?

When the uncertainty set is TV

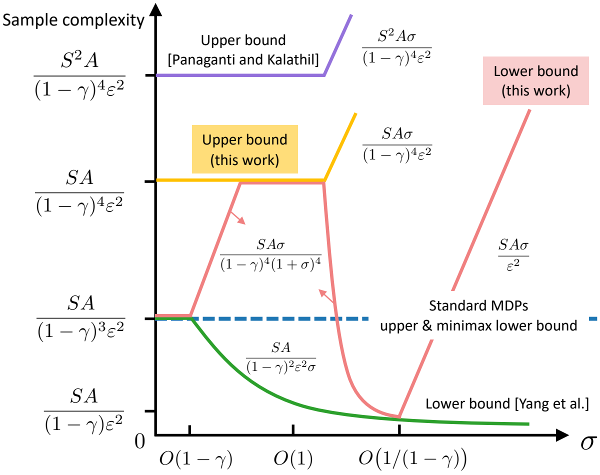


When the uncertainty set is TV

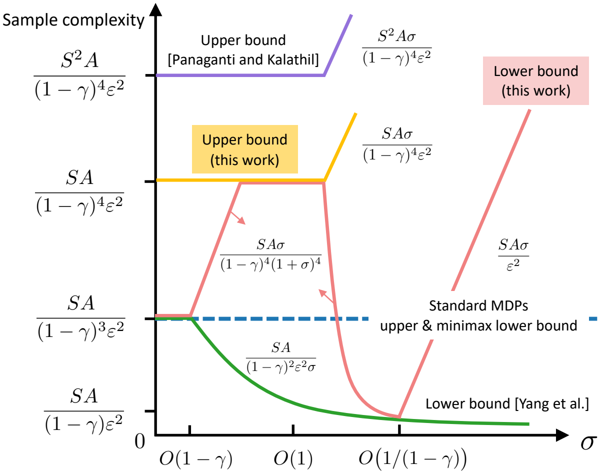


RMDPs are **easier** to learn than standard MDPs.

When the uncertainty set is χ^2 divergence



When the uncertainty set is χ^2 divergence



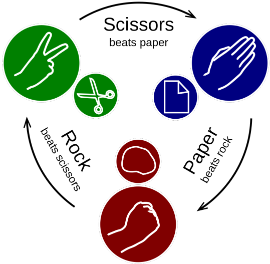
RMDPs can be **harder** to learn than standard MDPs.

Outline (Part 2)

Four variants of our basics settings to illustrate the approaches so far:

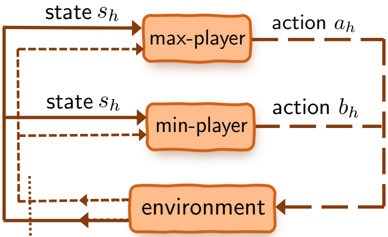
- Offline / batch RL
- RL with Markovian samples
- Robust RL
- Multi-agent RL

Background: two-player zero-sum Markov games



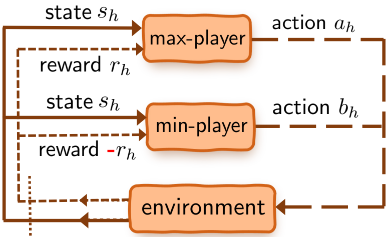
			
	0	-1	1
	1	0	-1
	-1	1	0

Two-player zero-sum Markov games



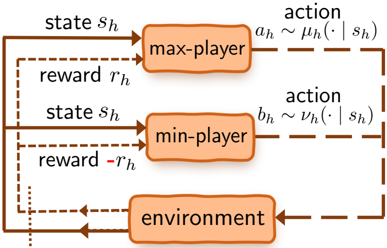
- $\mathcal{S} = [S]$: state space
- H: horizon
- $\mathcal{A} = [A]$: action space of max-player
- $\mathcal{B} = [B]$: action space of min-player

Two-player zero-sum Markov games



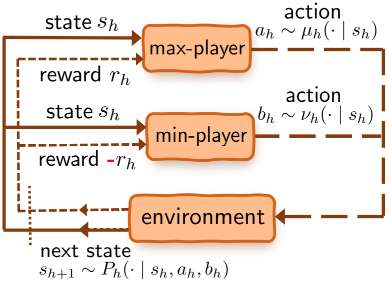
- $\mathcal{S} = [S]$: state space
- H: horizon
- $\mathcal{A} = [A]$: action space of max-player
- $\mathcal{B} = [B]$: action space of min-player
- immediate reward: max-player $r(s, a, b) \in [0, 1]$
  min-player $-r(s, a, b)$

Two-player zero-sum Markov games



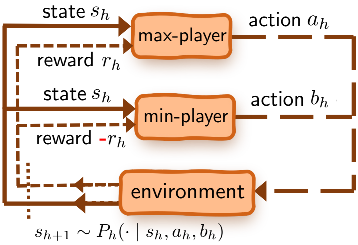
- $\mathcal{S} = [S]$: state space
- $\mathcal{A} = [A]$: action space of max-player
- H: horizon
- $\mathcal{B} = [B]$: action space of min-player
- immediate reward: max-player $r(s, a, b) \in [0, 1]$
  min-player $-r(s, a, b)$
- $\mu : \mathcal{S} \times [H] \to \Delta(\mathcal{A})$: policy of max-player
  $\nu : \mathcal{S} \times [H] \to \Delta(\mathcal{B})$: policy of min-player

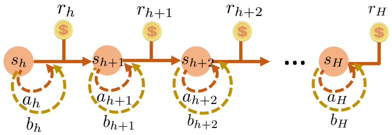
Two-player zero-sum Markov games



- $\mathcal{S} = [S]$: state space
- H: horizon
- $\mathcal{A} = [A]$: action space of max-player
- $\mathcal{B} = [B]$: action space of min-player
- immediate reward: max-player $r(s,a,b) \in [0,1]$
  min-player $-r(s,a,b)$
- $\mu : \mathcal{S} \times [H] \to \Delta(\mathcal{A})$: policy of max-player
  $\nu : \mathcal{S} \times [H] \to \Delta(\mathcal{B})$: policy of min-player
- $P_h(\cdot \mid s,a,b)$: unknown transition probabilities

Value function & Q-function

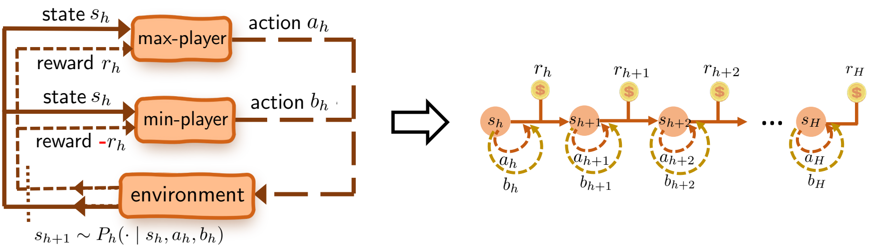




Value function of policy pair (μ, ν):

$$V_1^{\mu,\nu}(s) := \mathbb{E}\left[\sum_{t=1}^{H} r(s_t, a_t, b_t) \,\middle|\, s_1 = s\right]$$

Value function & Q-function

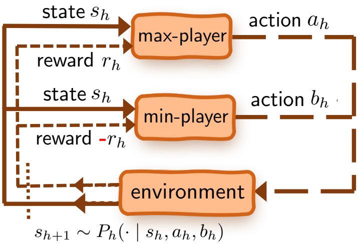


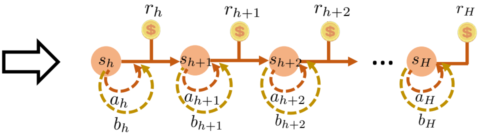
Value function of policy pair (μ, ν):

$$V_1^{\mu,\nu}(s) := \mathbb{E}\left[\sum_{t=1}^{H} r(s_t, a_t, b_t) \,\middle|\, s_1 = s\right]$$

- $(a_1, b_1, s_2, \cdots)$: generated when max-player and min-player execute policies μ and ν *independently (i.e., no coordination)*

Value function & Q-function





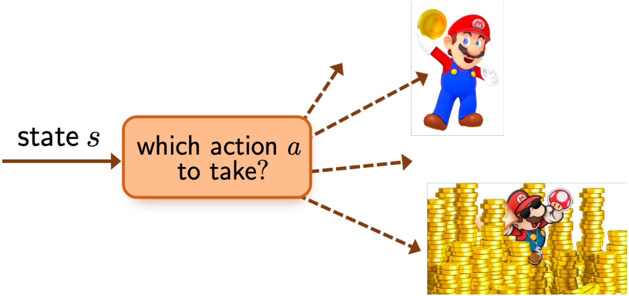
Value function and **Q function** of policy pair (μ, ν):

$$V_1^{\mu,\nu}(s) := \mathbb{E}\left[\sum_{t=1}^{H} r(s_t, a_t, b_t) \,\middle|\, s_1 = s\right]$$

$$Q_1^{\mu,\nu}(s, a, b) := \mathbb{E}\left[\sum_{t=1}^{H} r(s_t, a_t, b_t) \,\middle|\, s_1 = s, a_1 = a, b_1 = b\right]$$

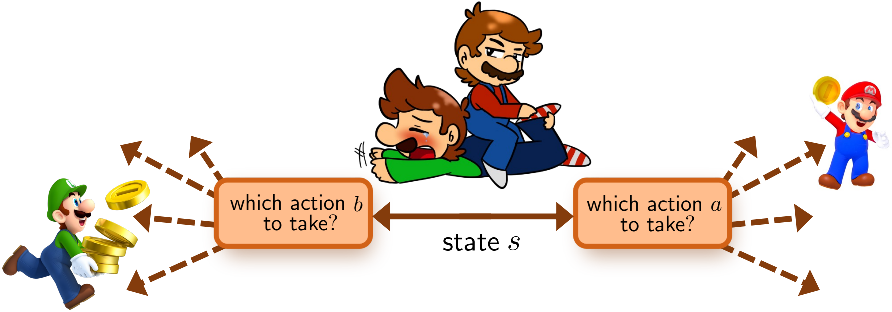
- $(a_1, b_1, s_2, \cdots)$: generated when max-player and min-player execute policies μ and ν *independently (i.e., no coordination)*

Optimal policy?



- Each agent seeks **optimal policy** maximizing her own value

Optimal policy?



- Each agent seeks **optimal policy** maximizing her own value
- But two agents have conflicting goals . . .

Compromise: Nash equilibrium (NE)

John von Neumann

John Nash

An NE policy pair $(\mu^\star, \nu^\star)$ obeys

$$\max_{\mu} V^{\mu,\nu^\star} = V^{\mu^\star,\nu^\star} = \min_{\nu} V^{\mu^\star,\nu}$$

Compromise: Nash equilibrium (NE)

John von Neumann

John Nash

An NE policy pair $(\mu^\star, \nu^\star)$ obeys

$$\max_{\mu} V^{\mu,\nu^\star} = V^{\mu^\star,\nu^\star} = \min_{\nu} V^{\mu^\star,\nu}$$

- no unilateral deviation is beneficial

Compromise: Nash equilibrium (NE)

John von Neumann

John Nash

An NE policy pair $(\mu^\star, \nu^\star)$ obeys

$$\max_{\mu} V^{\mu,\nu^\star} = V^{\mu^\star,\nu^\star} = \min_{\nu} V^{\mu^\star,\nu}$$

- no unilateral deviation is beneficial
- no coordination between two agents (they act *independently*)

Compromise: Nash equilibrium (NE)

John von Neumann

John Nash

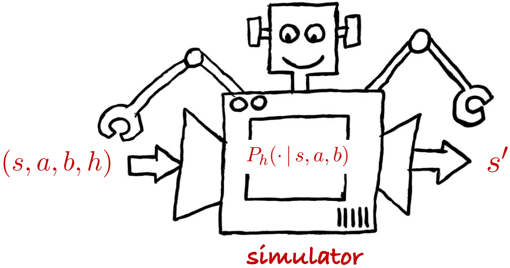
An ε-NE policy pair $(\widehat{\mu}, \widehat{\nu})$ obeys

$$\max_{\mu} V^{\mu,\widehat{\nu}} - \varepsilon \leq V^{\widehat{\mu},\widehat{\nu}} \leq \min_{\nu} V^{\widehat{\mu},\nu} + \varepsilon$$

- no unilateral deviation is beneficial
- no coordination between two agents (they act *independently*)

Sampling mechanism: a generative model / simulator

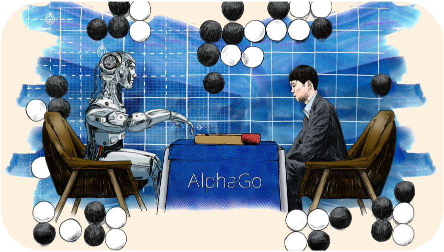
— *Kearns, Singh '99*

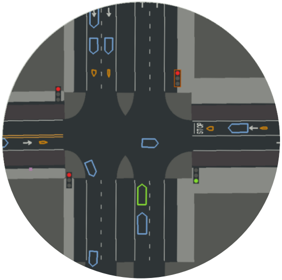


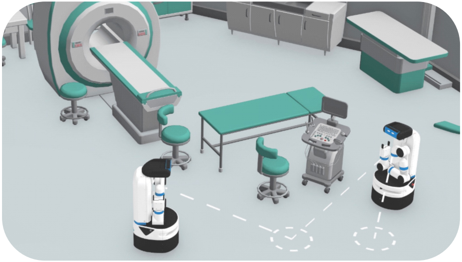
One can query generative model w/ state-action-step tuple (s, a, b, h), and obtain $s' \overset{\text{ind.}}{\sim} P_h(s' \mid s, a, b)$

Question: *how many samples are sufficient to learn an ε-Nash policy pair?*

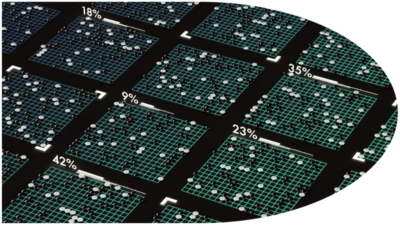
Multi-agent reinforcement learning (MARL)







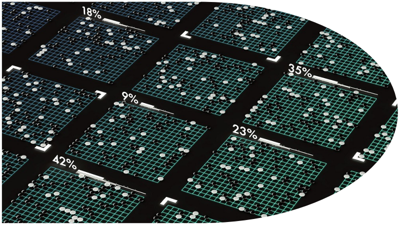
Challenges



In MARL, agents learn by probing the (shared) environment

- unknown or changing environment
- delayed feedback
- explosion of dimensionality

Challenges

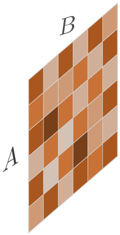


In MARL, agents learn by probing the (shared) environment

- unknown or changing environment
- delayed feedback
- explosion of dimensionality
- **curse of multiple agents**

Model-based approach w/ non-adaptive sampling

— *Zhang, Kakade, Başar, Yang '20*

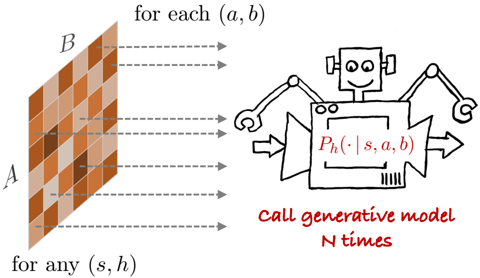


for any (s, h)

1. for each (s, a, b, h), call generative models N times

Model-based approach w/ non-adaptive sampling

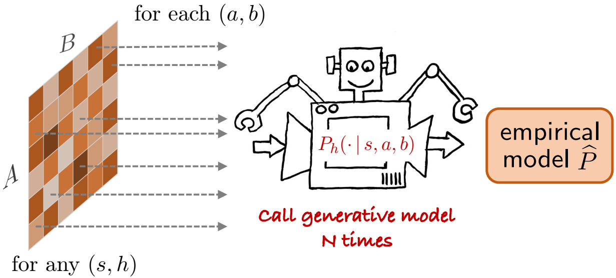
— *Zhang, Kakade, Başar, Yang '20*



1. for each (s, a, b, h), call generative models N times

Model-based approach w/ non-adaptive sampling

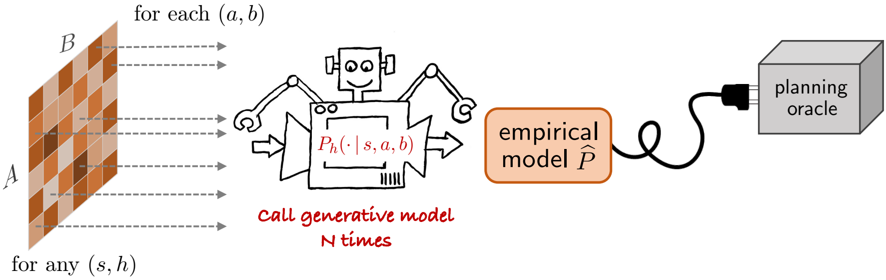
— *Zhang, Kakade, Başar, Yang '20*



1. for each (s, a, b, h), call generative models N times
2. build empirical model $\widehat{P}$

Model-based approach w/ non-adaptive sampling

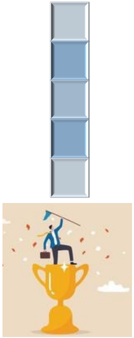
— Zhang, Kakade, Başar, Yang '20



1. for each (s, a, b, h), call generative models N times
2. build empirical model $\widehat{P}$, and run classical planning algorithms

sample complexity: $\frac{H^4 SAB}{\varepsilon^2}$

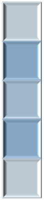
Curse of multiple agents

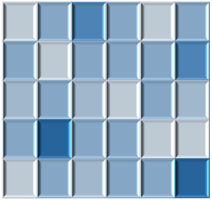


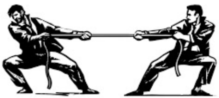
1 player: A

Let's look at the size of joint action space . . .

Curse of multiple agents





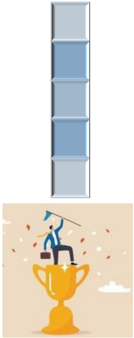


1 player: A

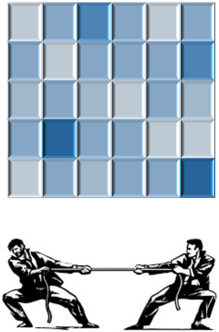
2 players: AB

Let's look at the size of joint action space ...

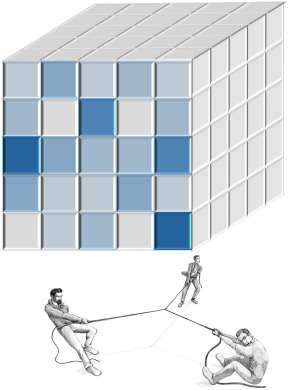
Curse of multiple agents



1 player: A



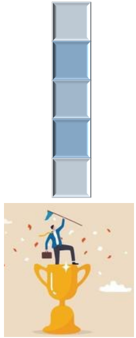
2 players: AB



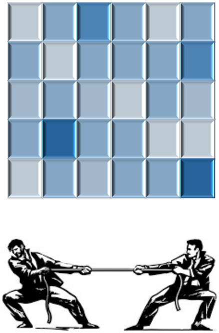
3 players: $A_1 A_2 A_3$

Let's look at the size of joint action space . . .

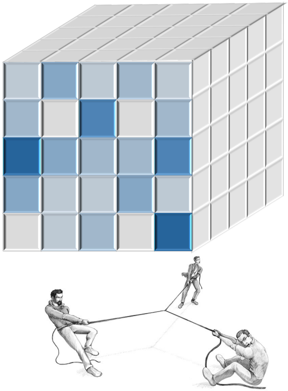
Curse of multiple agents



1 player: A



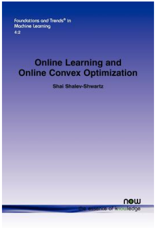
2 players: AB



3 players: $A_1 A_2 A_3$

The number of joint actions blows up geometrically in # players!

Breaking curse of multi-agents?

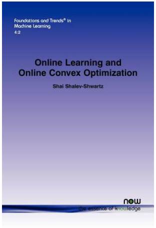


— *Song, Mei, Bai '21, Jin, Liu, Wang, Yu '21, . . .*

V-learning: overcomes curse of multi-agents in *online* RL

- estimate V-function only (much lower-dimensional than Q)

Breaking curse of multi-agents?

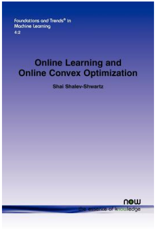


— *Song, Mei, Bai '21, Jin, Liu, Wang, Yu '21, . . .*

V-learning: overcomes curse of multi-agents in *online* RL

- estimate V-function only (much lower-dimensional than Q)
- *adaptive sampling:* take sample based on current policy iterates

Breaking curse of multi-agents?

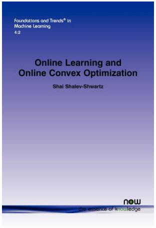


— *Song, Mei, Bai '21, Jin, Liu, Wang, Yu '21, . . .*

V-learning: overcomes curse of multi-agents in *online* RL

- estimate V-function only (much lower-dimensional than Q)
- *adaptive sampling:* take sample based on current policy iterates
- *adversarial learning subroutine*: Follow-the-Regularized-Leader

Breaking curse of multi-agents?

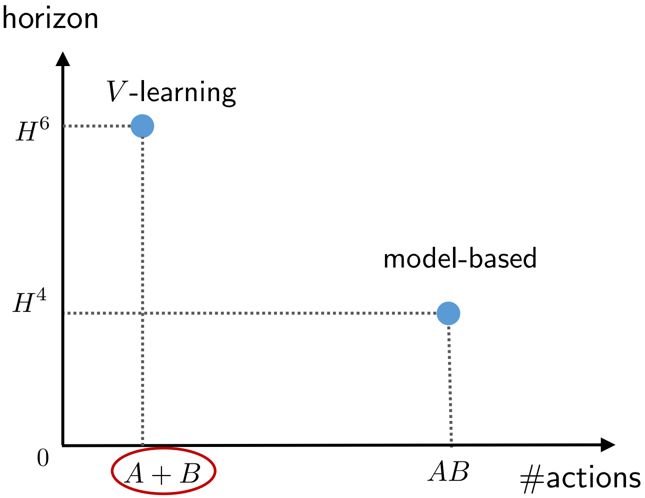


— *Song, Mei, Bai '21, Jin, Liu, Wang, Yu '21, . . .*

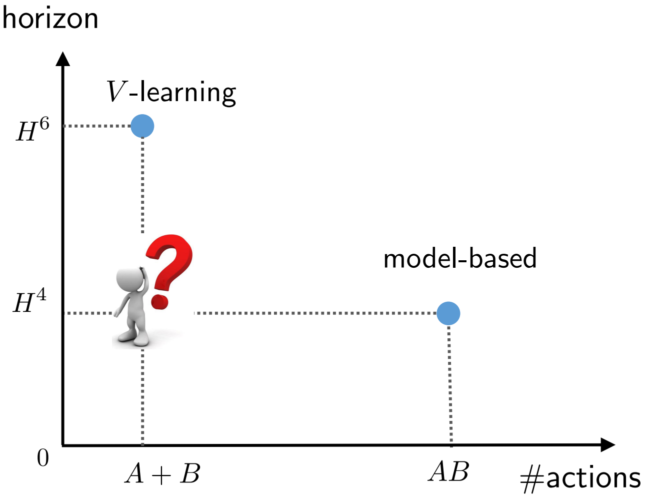
V-learning: overcomes curse of multi-agents in *online* RL

- estimate V-function only (much lower-dimensional than Q)
- *adaptive sampling:* take sample based on current policy iterates
- *adversarial learning subroutine*: Follow-the-Regularized-Leader

sample complexity: $\frac{H^6 S(A+B)}{\varepsilon^2}$ samples or $\frac{H^5 S(A+B)}{\varepsilon^2}$ episodes



horizon
V-learning
H^6
model-based
H^4
0
$A + B$
AB
#actions



Can we simultaneously overcome
curse of multi-agents & barrier of long horizon?

Our algorithm

Key ingredients:

- for each player, estimate only **one-sided objects**
  - e.g. $Q(s, a)$ as opposed to $Q(s, a, b)$

Our algorithm

Key ingredients:

- for each player, estimate only **one-sided objects**
  - e.g. $Q(s, a)$ as opposed to $Q(s, a, b)$
- **adaptive sampling**
  - sampling based on current policy iterates

Our algorithm

Key ingredients:

- for each player, estimate only **one-sided objects**
  - e.g. $Q(s,a)$ as opposed to $Q(s,a,b)$
- **adaptive sampling**
  - sampling based on current policy iterates
- **adversarial learning subroutine** for policy updates
  - e.g. Follow-the-Regularized-Leader (FTRL)

Our algorithm

Key ingredients:

- for each player, estimate only **one-sided objects**
  - e.g. $Q(s,a)$ as opposed to $Q(s,a,b)$
- **adaptive sampling**
  - sampling based on current policy iterates
- **adversarial learning subroutine** for policy updates
  - e.g. Follow-the-Regularized-Leader (FTRL)
- **optimism principle** in value estimation
  - upper confidence bounds (UCB)

Main result (two-player zero-sum Markov games)

Theorem (Li, Chi, Wei, Chen '22)

For any $0 < \varepsilon \leq H$, the policy pair $(\widehat{\mu}, \widehat{\nu})$ returned by the proposed algorithm is ε-Nash, with sample complexity at most

$$\widetilde{O}\left(\frac{H^4 S(A+B)}{\varepsilon^2}\right)$$

Main result (two-player zero-sum Markov games)

Theorem (Li, Chi, Wei, Chen '22)

For any $0 < \varepsilon \leq H$, the policy pair $(\widehat{\mu}, \widehat{\nu})$ returned by the proposed algorithm is ε-Nash, with sample complexity at most

$$\widetilde{O}\left(\frac{H^4 S(A+B)}{\varepsilon^2}\right)$$

- **minimax lower bound:** $\widetilde{\Omega}\big(\frac{H^4 S(A+B)}{\varepsilon^2}\big)$
- breaks curse of multi-agents & long-horizon barrier at once!

Main result (two-player zero-sum Markov games)

Theorem (Li, Chi, Wei, Chen '22)

For any $0 < \varepsilon \leq H$, the policy pair $(\widehat{\mu}, \widehat{\nu})$ returned by the proposed algorithm is ε-Nash, with sample complexity at most

$$\widetilde{O}\left(\frac{H^4 S(A+B)}{\varepsilon^2}\right)$$

- **minimax lower bound:** $\widetilde{\Omega}\left(\frac{H^4 S(A+B)}{\varepsilon^2}\right)$
- breaks curse of multi-agents & long-horizon barrier at once!
- full ε-range (no burn-in cost)

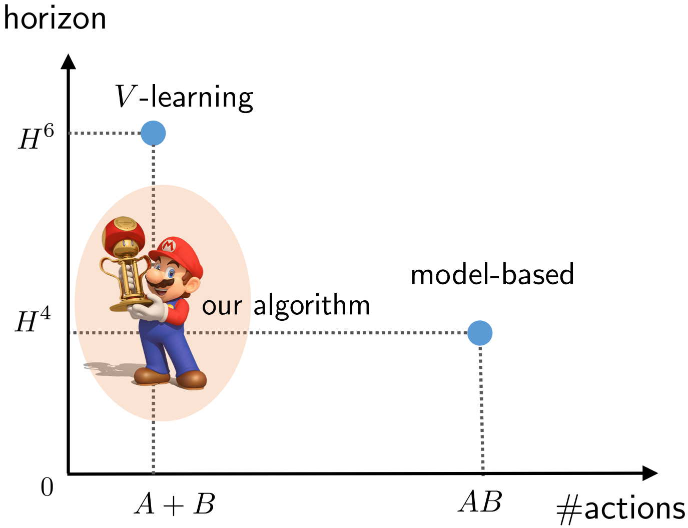
Main result (two-player zero-sum Markov games)

Theorem (Li, Chi, Wei, Chen '22)

For any $0 < \varepsilon \leq H$, the policy pair $(\widehat{\mu}, \widehat{\nu})$ returned by the proposed algorithm is ε-Nash, with sample complexity at most

$$\widetilde{O}\left(\frac{H^4 S(A+B)}{\varepsilon^2}\right)$$

- **minimax lower bound:** $\widetilde{\Omega}\left(\frac{H^4 S(A+B)}{\varepsilon^2}\right)$
- breaks curse of multi-agents & long-horizon barrier at once!
- full ε-range (no burn-in cost)
- other features: Markov policy, decentralized, ...


horizon
V-learning
H^6
model-based
our algorithm
H^4
0
$A + B$
AB
#actions

Extension: m-player general-sum Markov games

Theorem (Li, Chi, Wei, Chen '22)

For any $0 < \varepsilon \leq H$, the joint policy $\widehat{\pi}$ returned by the proposed algorithm is ε-CCE, with sample complexity at most

$$\widetilde{O}\left(\frac{H^4 S \sum_i A_i}{\varepsilon^2}\right)$$

Extension: m-player general-sum Markov games

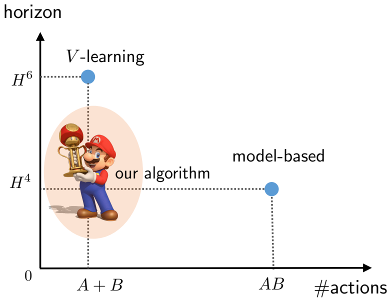
Theorem (Li, Chi, Wei, Chen '22)

For any $0 < \varepsilon \leq H$, the joint policy $\widehat{\pi}$ returned by the proposed algorithm is ε-CCE, with sample complexity at most

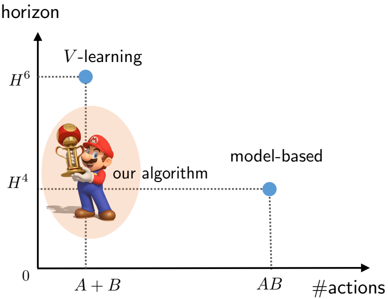
$$\widetilde{O}\left(\frac{H^4 S \sum_i A_i}{\varepsilon^2}\right)$$

- **minimax lower bound:** $\widetilde{\Omega}\big(\frac{H^4 S \max_i A_i}{\varepsilon^2}\big)$
- near-optimal when number of players m is fixed

Overcomes curse of multi-agents and long-horizon barrier simultaneously in the presence of generative model!



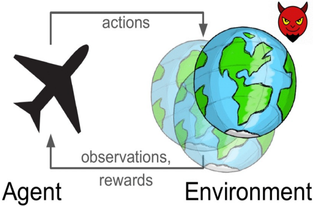
Overcomes curse of multi-agents and long-horizon barrier simultaneously in the presence of generative model!

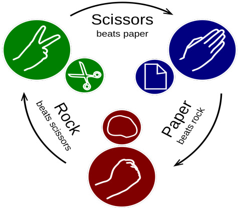


Future directions:

- optimal sample complexity for CCE when # players is large
- optimal sample complexity for online RL

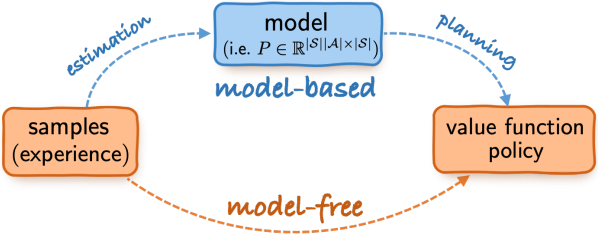
Summary of this part





Four variants of our basics settings:
offline RL / RL with Markovian samples / robust RL / multi-agent RL

Recall: three approaches



Model-based approach ("plug-in")

- build an empirical estimate $\widehat{P}$ for P
- planning based on the empirical $\widehat{P}$

Value-based approach

— learning w/o estimating the model explicitly

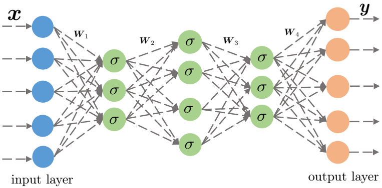
Policy-based approach

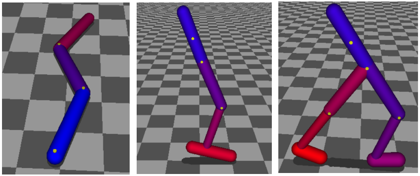
— optimization in the space of policies

Policy optimization in practice

$$\text{maximize}_\theta \quad \text{value}(\text{policy}(\theta))$$

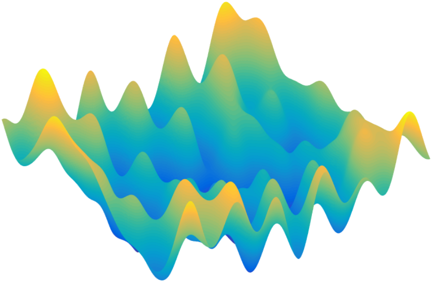
- directly optimize the policy, which is the quantity of interest
- allow flexible differentiable parameterizations of the policy
- work with both continuous and discrete problems





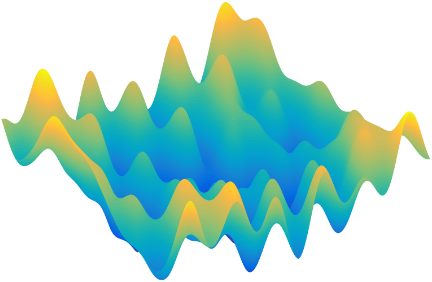
Theoretical challenges: non-concavity

Little understanding on the global convergence of policy gradient methods until very recently, e.g. (Fazel et al., 2018; Bhandari and Russo, 2019; Agarwal et al., 2019; Mei et al. 2020), and many more.



Theoretical challenges: non-concavity

Little understanding on the global convergence of policy gradient methods until very recently, e.g. (Fazel et al., 2018; Bhandari and Russo, 2019; Agarwal et al., 2019; Mei et al. 2020), and many more.



Our goal:

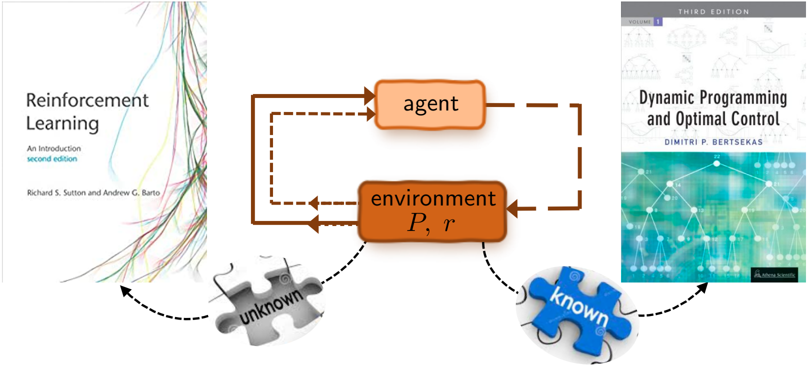
- understand finite-time convergence rates of popular heuristics
- design fast-convergent algorithms that scale for finding policies with desirable properties

Outline

- Backgrounds and basics
  - policy gradient method
- Convergence guarantees of single-agent policy optimization
  - (natural) policy gradient methods
  - finite-time rate of global convergence
  - entropy regularization and beyond
- Concluding remarks

Backgrounds: policy optimization in tabular Markov decision processes

Searching for the optimal policy



Goal: find the optimal policy $\pi^\star$ that maximize $V^\pi(s)$

- optimal value / Q function: $V^\star := V^{\pi^\star}$, $Q^\star := Q^{\pi^\star}$

Policy gradient methods

Given an initial state distribution $s \sim \rho$, find policy π such that

$$\text{maximize}_{\pi} \quad V^{\pi}(\rho) := \mathbb{E}_{s \sim \rho}\left[V^{\pi}(s)\right]$$

Policy gradient methods

Given an initial state distribution $s \sim \rho$, find policy π such that

$$\text{maximize}_{\pi} \quad V^{\pi}(\rho) := \mathbb{E}_{s\sim\rho}\left[V^{\pi}(s)\right]$$

$$\Downarrow$$

Parameterization:

$$\pi := \pi_\theta$$

Policy gradient methods

Given an initial state distribution $s \sim \rho$, find policy π such that

$$\text{maximize}_{\pi} \quad V^{\pi}(\rho) := \mathbb{E}_{s\sim\rho}\left[V^{\pi}(s)\right]$$

$$\Downarrow$$

Parameterization:

$$\pi := \pi_{\theta}$$

$$\text{maximize}_{\theta} \quad V^{\pi_\theta}(\rho) := \mathbb{E}_{s\sim\rho}\left[V^{\pi_\theta}(s)\right]$$

Policy gradient methods

Given an initial state distribution $s \sim \rho$, find policy π such that

$$\text{maximize}_{\pi} \quad V^{\pi}(\rho) := \mathbb{E}_{s\sim\rho}\left[V^{\pi}(s)\right]$$

$$\Downarrow$$

Parameterization:

$$\pi := \pi_\theta$$

$$\text{maximize}_{\theta} \quad V^{\pi_\theta}(\rho) := \mathbb{E}_{s\sim\rho}\left[V^{\pi_\theta}(s)\right]$$

Policy gradient method (Sutton et al., 2000)

For $t = 0, 1, \cdots$

$$\theta^{(t+1)} = \theta^{(t)} + \eta \nabla_\theta V^{\pi_\theta^{(t)}}(\rho)$$

where η *is the learning rate.*

Softmax policy gradient methods

Given an initial state distribution $s \sim \rho$, find policy π such that

$$\text{maximize}_{\pi} \quad V^{\pi}(\rho) := \mathbb{E}_{s\sim\rho}\left[V^{\pi}(s)\right]$$

$\Downarrow$ softmax parameterization:

$$\pi_\theta(a|s) \propto \exp(\theta(s,a))$$

$$\text{maximize}_{\theta} \quad V^{\pi_\theta}(\rho) := \mathbb{E}_{s\sim\rho}\left[V^{\pi_\theta}(s)\right]$$

Policy gradient method (Sutton et al., 2000)

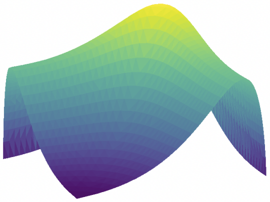
For $t = 0, 1, \cdots$

$$\theta^{(t+1)} = \theta^{(t)} + \eta \nabla_\theta V^{\pi_\theta^{(t)}}(\rho)$$

where η *is the learning rate.*

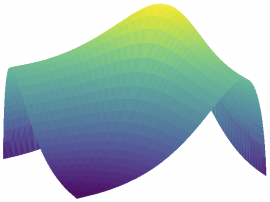
Finite-time global convergence guarantees

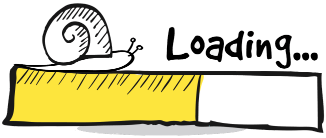
Global convergence of the PG method?



- (Agarwal et al., 2019) showed that softmax PG converges asymptotically to the global optimal policy.

Global convergence of the PG method?

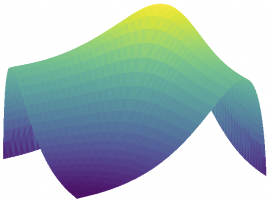


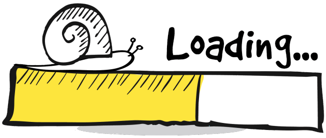


- (Agarwal et al., 2019) showed that softmax PG converges asymptotically to the global optimal policy.
- (Mei et al., 2020) Softmax PG converges to global opt in

$$O\left(\frac{1}{\varepsilon}\right) \text{ iterations}$$

Global convergence of the PG method?

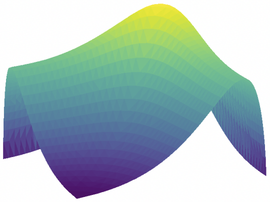


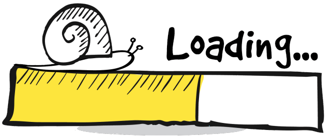


- (Agarwal et al., 2019) showed that softmax PG converges asymptotically to the global optimal policy.
- (Mei et al., 2020) Softmax PG converges to global opt in

$$c\Big(|\mathcal{S}|, |\mathcal{A}|, \frac{1}{1-\gamma}, \cdots\Big)\, O\Big(\frac{1}{\varepsilon}\Big) \text{ iterations}$$

Global convergence of the PG method?





- (Agarwal et al., 2019) showed that softmax PG converges asymptotically to the global optimal policy.
- (Mei et al., 2020) Softmax PG converges to global opt in

$$c\left(|\mathcal{S}|, |\mathcal{A}|, \frac{1}{1-\gamma}, \cdots\right) O\left(\frac{1}{\varepsilon}\right) \text{ iterations}$$

Is the rate of PG good, bad or ugly?

A negative message

Theorem (Li, Wei, Chi, Chen, 2021)

There exists an MDP s.t. it takes softmax PG at least

$$\frac{1}{\eta}\,|\mathcal{S}|^{2^{\Theta(\frac{1}{1-\gamma})}} \textit{ iterations}$$

to achieve $\|V^{(t)} - V^{\star}\|_{\infty} \leq 0.15$.

A negative message

Theorem (Li, Wei, Chi, Chen, 2021)

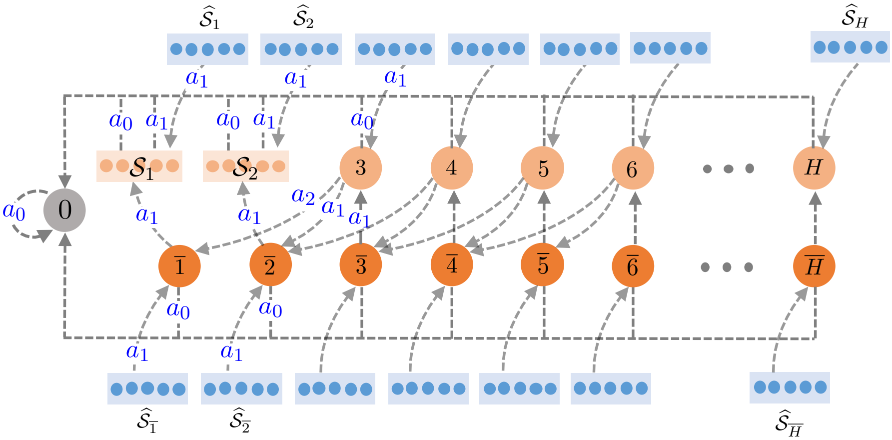
There exists an MDP s.t. it takes softmax PG at least

$$\frac{1}{\eta}|\mathcal{S}|^{2^{\Theta(\frac{1}{1-\gamma})}} \textit{ iterations}$$

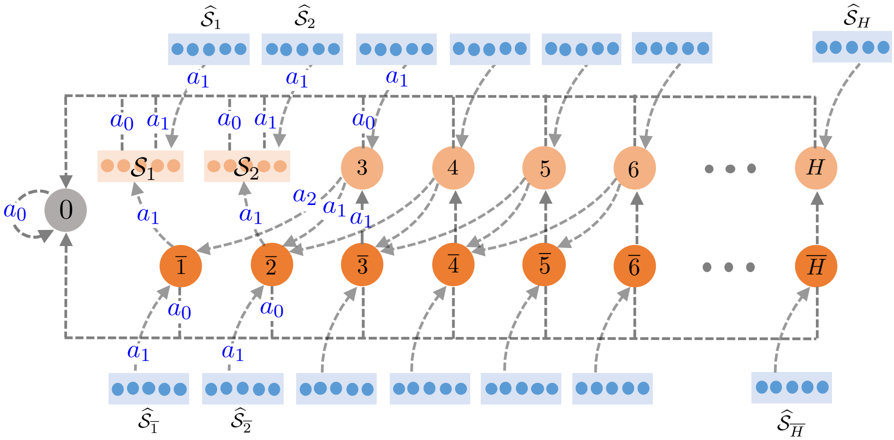
to achieve $\|V^{(t)} - V^{\star}\|_{\infty} \leq 0.15$.

- Softmax PG can take (super)-exponential time to converge (in problems w/ large state space & long effective horizon)!
- Also hold for average sub-opt gap $\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\big[V^{(t)}(s) - V^{\star}(s)\big]$.

MDP construction for our lower bound

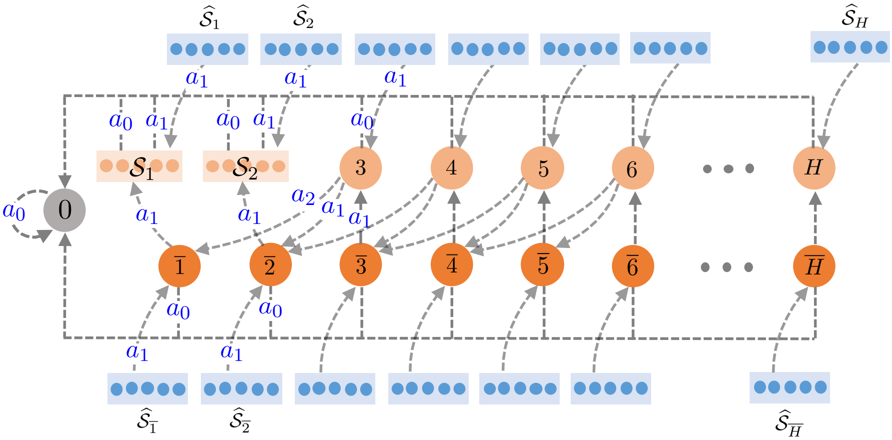


MDP construction for our lower bound



Key ingredients: for $3 \leq s \leq H \asymp \frac{1}{1-\gamma}$,

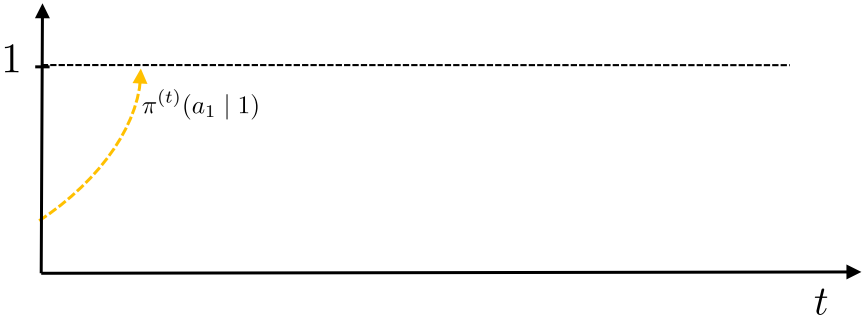
MDP construction for our lower bound



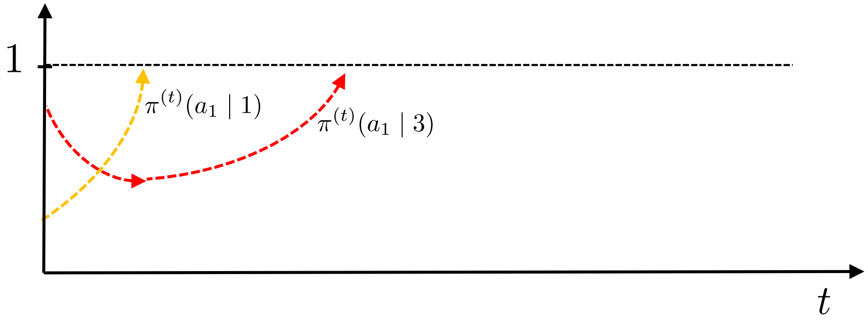
Key ingredients: for $3 \leq s \leq H \asymp \frac{1}{1-\gamma}$,

- $\pi^{(t)}(a_{\mathsf{opt}} \mid s)$ keeps decreasing until $\pi^{(t)}(a_{\mathsf{opt}} \mid s-2) \approx 1$

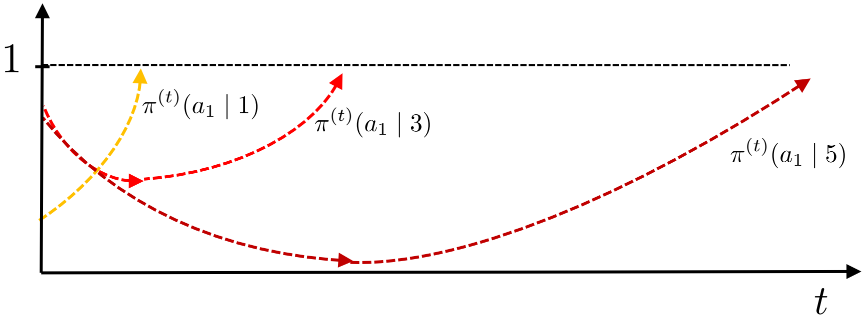
What is happening in our constructed MDP?



What is happening in our constructed MDP?

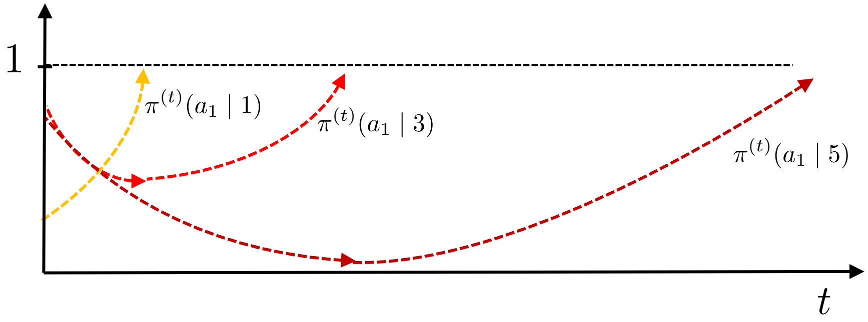


What is happening in our constructed MDP?



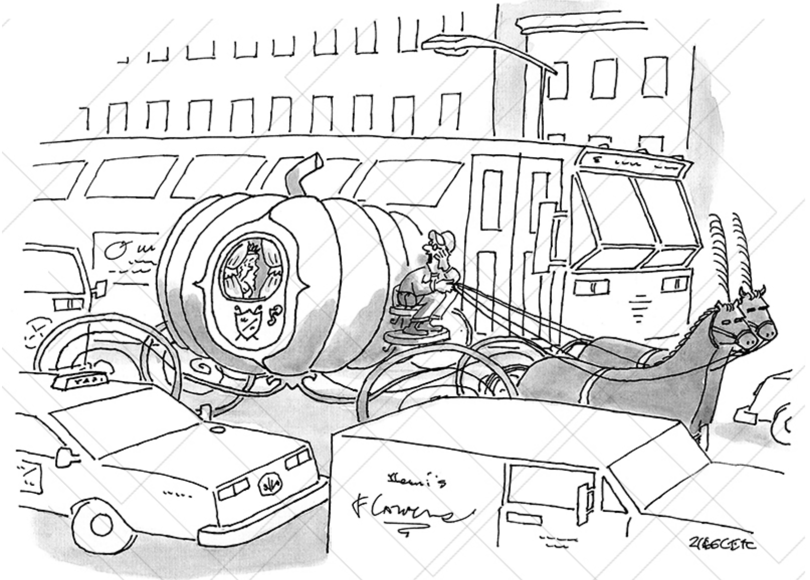
Convergence time for state s grows geometrically as s increases

What is happening in our constructed MDP?



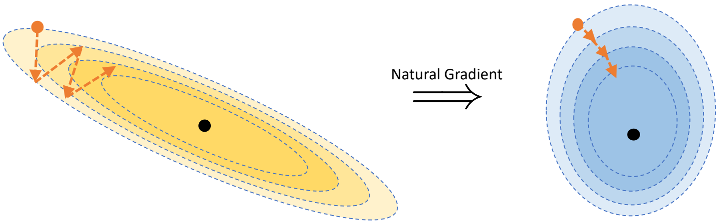
Convergence time for state s grows geometrically as s increases

$$\text{convergence-time}(s) \gtrsim \big(\text{convergence-time}(s-2)\big)^{1.5}$$



"Seriously, lady, at this hour you'd make a lot better time taking the subway."

Booster #1: natural policy gradient



Natural policy gradient (NPG) method (Kakade, 2002)

For $t = 0, 1, \cdots$

$$\theta^{(t+1)} = \theta^{(t)} + \eta (\mathcal{F}_\rho^\theta)^\dagger \nabla_\theta V^{\pi_\theta^{(t)}}(\rho)$$

where η *is the learning rate and* $\mathcal{F}_\rho^\theta$ *is the Fisher information matrix:*

$$\mathcal{F}_\rho^\theta := \mathbb{E}\left[\big(\nabla_\theta \log \pi_\theta(a|s)\big)\big(\nabla_\theta \log \pi_\theta(a|s)\big)^\top\right].$$

Connection with TRPO/PPO

TRPO/PPO (Schulman et al., 2015; 2017) are popular heuristics in training RL algorithms, with **KL regularization**

$$\mathsf{KL}(\pi_\theta^{(t)} \| \pi_\theta) \approx \frac{1}{2}(\theta - \theta^{(t)})^\top \mathcal{F}_\rho^\theta (\theta - \theta^{(t)})$$

via constrained or proximal terms:

$$\begin{aligned} \theta^{(t+1)} &= \operatorname*{argmax}_\theta V^{\pi_\theta^{(t)}}(\rho) + (\theta - \theta^{(t)})^\top \nabla_\theta V^{\pi_\theta^{(t)}}(\rho) - \eta \mathsf{KL}(\pi_\theta^{(t)} \| \pi_\theta) \\ &\approx \theta^{(t)} + \eta (\mathcal{F}_\rho^\theta)^\dagger \nabla_\theta V^{\pi_\theta^{(t)}}(\rho), \end{aligned}$$

leading to exactly NPG!

Connection with TRPO/PPO

TRPO/PPO (Schulman et al., 2015; 2017) are popular heuristics in training RL algorithms, with **KL regularization**

$$\mathsf{KL}(\pi_\theta^{(t)} \| \pi_\theta) \approx \frac{1}{2}(\theta - \theta^{(t)})^\top \mathcal{F}_\rho^\theta (\theta - \theta^{(t)})$$

via constrained or proximal terms:

$$\begin{aligned} \theta^{(t+1)} &= \arg\max_\theta V^{\pi_\theta^{(t)}}(\rho) + (\theta - \theta^{(t)})^\top \nabla_\theta V^{\pi_\theta^{(t)}}(\rho) - \eta \mathsf{KL}(\pi_\theta^{(t)} \| \pi_\theta) \\ &\approx \theta^{(t)} + \eta (\mathcal{F}_\rho^\theta)^\dagger \nabla_\theta V^{\pi_\theta^{(t)}}(\rho), \end{aligned}$$

leading to exactly NPG!

NPG $\approx$ TRPO/PPO!

NPG in the tabular setting

Natural policy gradient (NPG) method (Tabular setting)

For $t = 0, 1, \cdots$, NPG updates the policy via

$$\pi^{(t+1)}(\cdot|s) \propto \underbrace{\pi^{(t)}(\cdot|s)}_{\text{current policy}} \underbrace{\exp\left(\frac{\eta Q^{(t)}(s,\cdot)}{1-\gamma}\right)}_{\text{soft greedy}}$$

where $Q^{(t)} := Q^{\pi^{(t)}}$ is the Q-function of $\pi^{(t)}$, and $\eta > 0$.

- invariant with the choice of ρ
- Reduces to policy iteration (PI) when $\eta = \infty$.

Global convergence of NPG

Theorem (Agarwal et al., 2019)

Set $\pi^{(0)}$ *as a uniform policy. For all* $t \geq 0$, *we have*

$$V^{(t)}(\rho) \geq V^{\star}(\rho) - \left(\frac{\log |\mathcal{A}|}{\eta} + \frac{1}{(1-\gamma)^2} \right) \frac{1}{t}.$$

Global convergence of NPG

Theorem (Agarwal et al., 2019)

Set $\pi^{(0)}$ *as a uniform policy. For all* $t \geq 0$, *we have*

$$V^{(t)}(\rho) \geq V^{\star}(\rho) - \left(\frac{\log |\mathcal{A}|}{\eta} + \frac{1}{(1-\gamma)^2} \right) \frac{1}{t}.$$

Implication: set $\eta \geq (1-\gamma)^2 \log |\mathcal{A}|$, we find an ϵ-optimal policy within at most

$$\frac{2}{(1-\gamma)^2 \epsilon} \quad \text{iterations.}$$

Global convergence of NPG

Theorem (Agarwal et al., 2019)

Set $\pi^{(0)}$ as a uniform policy. For all $t \geq 0$, we have

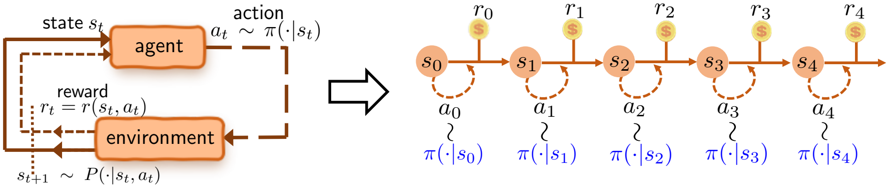
$$V^{(t)}(\rho) \geq V^{\star}(\rho) - \left(\frac{\log |\mathcal{A}|}{\eta} + \frac{1}{(1-\gamma)^2} \right) \frac{1}{t}.$$

Implication: set $\eta \geq (1-\gamma)^2 \log |\mathcal{A}|$, we find an ϵ-optimal policy within at most

$$\frac{2}{(1-\gamma)^2 \epsilon} \quad \text{iterations.}$$

Global convergence at a sublinear rate independent of $|\mathcal{S}|$, $|\mathcal{A}|$!

Booster #2: entropy regularization

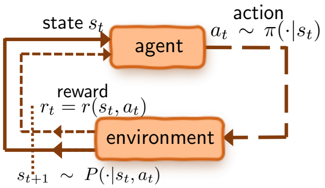


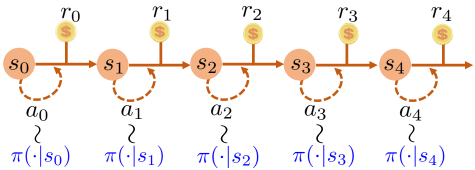
To encourage exploration, promote the stochasticity of the policy using the **"soft"** value function (Williams and Peng, 1991):

$$\forall s \in \mathcal{S}: \qquad V_\tau^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \Big(r_t + \tau \mathcal{H}(\pi(\cdot|s_t))\Big) \,\Big|\, s_0 = s\right]$$

where $\mathcal{H}$ is the Shannon entropy, and $\tau \geq 0$ is the reg. parameter.

Booster #2: entropy regularization





To encourage exploration, promote the stochasticity of the policy using the **"soft"** value function (Williams and Peng, 1991):

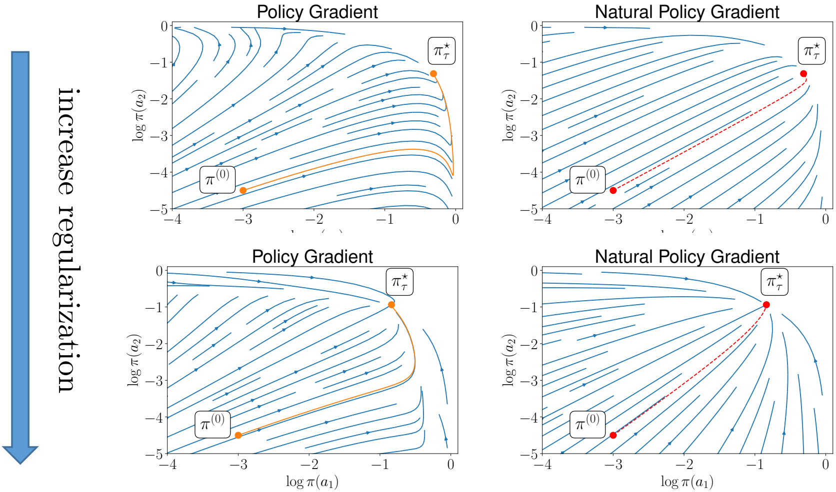
$$\forall s \in \mathcal{S}: \qquad V_\tau^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \Big(r_t + \tau \mathcal{H}(\pi(\cdot|s_t))\Big) \,\Big|\, s_0 = s\right]$$

where $\mathcal{H}$ is the Shannon entropy, and $\tau \geq 0$ is the reg. parameter.

$$\text{maximize}_\theta \quad V_\tau^{\pi_\theta}(\rho) := \mathbb{E}_{s\sim\rho}\left[V_\tau^{\pi_\theta}(s)\right]$$

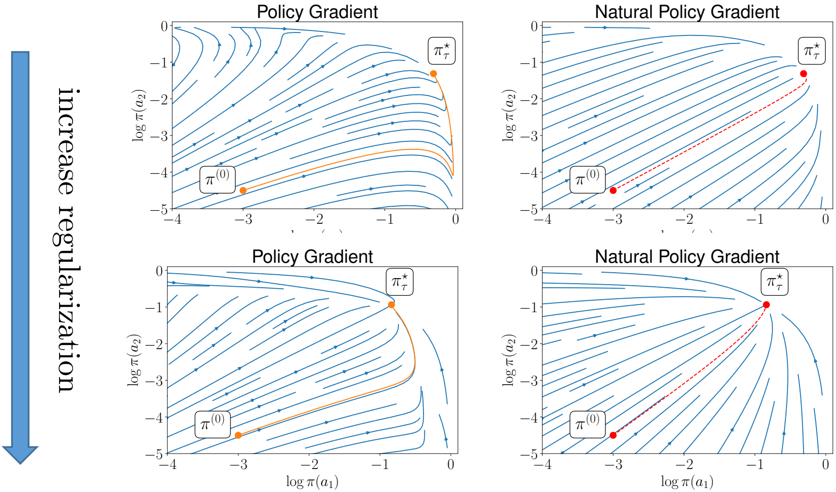
Entropy-regularized natural gradient helps!

Toy example: a bandit with 3 arms of rewards 1, 0.9 and 0.1.

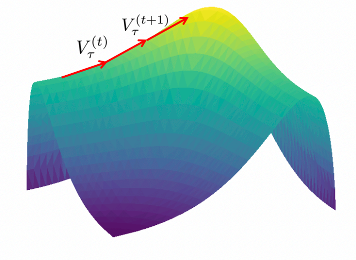


Entropy-regularized natural gradient helps!

Toy example: a bandit with 3 arms of rewards 1, 0.9 and 0.1.

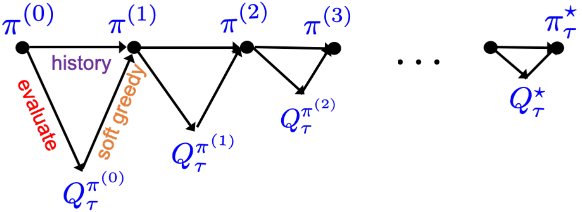


Can we justify the efficacy of entropy-regularized NPG?



How to characterize the efficiency of entropy-regularized NPG in tabular settings?

Entropy-regularized NPG in the tabular setting



Entropy-regularized NPG (Tabular setting)

For $t = 0, 1, \cdots$, *the policy is updated via*

$$\pi^{(t+1)}(\cdot|s) \propto \underbrace{\pi^{(t)}(\cdot|s)}_{\textit{current policy}}{}^{1-\frac{\eta\tau}{1-\gamma}} \underbrace{\exp(Q_\tau^{(t)}(s,\cdot)/\tau)}_{\textit{soft greedy}}{}^{\frac{\eta\tau}{1-\gamma}}$$

where $Q_\tau^{(t)} := Q_\tau^{\pi^{(t)}}$ *is the soft Q-function of* $\pi^{(t)}$, *and* $0 < \eta \le \frac{1-\gamma}{\tau}$.

- invariant with the choice of ρ
- Reduces to soft policy iteration (SPI) when $\eta = \frac{1-\gamma}{\tau}$.

Linear convergence with exact gradient

Exact oracle: perfect evaluation of $Q_\tau^{\pi^{(t)}}$ given $\pi^{(t)}$;

Linear convergence with exact gradient

Exact oracle: perfect evaluation of $Q_\tau^{\pi^{(t)}}$ given $\pi^{(t)}$;

Theorem (Cen, Cheng, Chen, Wei, Chi, 2020)

For any learning rate $0 < \eta \leq (1-\gamma)/\tau$, the entropy-regularized NPG updates satisfy

- **Linear convergence of soft Q-functions:**

$$\|Q_\tau^\star - Q_\tau^{(t+1)}\|_\infty \leq C_1 \gamma \, (1 - \eta\tau)^t$$

for all $t \geq 0$, where $Q_\tau^\star$ is the optimal soft Q-function, and

$$C_1 = \|Q_\tau^\star - Q_\tau^{(0)}\|_\infty + 2\tau \left(1 - \frac{\eta\tau}{1-\gamma}\right) \|\log \pi_\tau^\star - \log \pi^{(0)}\|_\infty .$$

Implications

To reach $\|Q_\tau^\star - Q_\tau^{(t+1)}\|_\infty \le \epsilon$, the iteration complexity is at most

- **General learning rates ($0 < \eta < \frac{1-\gamma}{\tau}$):**

$$\frac{1}{\eta\tau}\log\left(\frac{C_1\gamma}{\epsilon}\right)$$

- **Soft policy iteration ($\eta = \frac{1-\gamma}{\tau}$):**

$$\frac{1}{1-\gamma}\log\left(\frac{\|Q_\tau^\star - Q_\tau^{(0)}\|_\infty\gamma}{\epsilon}\right)$$

Implications

To reach $\|Q_\tau^\star - Q_\tau^{(t+1)}\|_\infty \le \epsilon$, the iteration complexity is at most

- **General learning rates** ($0 < \eta < \frac{1-\gamma}{\tau}$)**:**

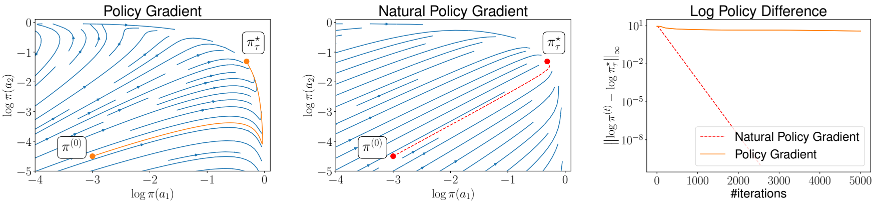
$$\frac{1}{\eta\tau}\log\left(\frac{C_1\gamma}{\epsilon}\right)$$

- **Soft policy iteration** ($\eta = \frac{1-\gamma}{\tau}$)**:**

$$\frac{1}{1-\gamma}\log\left(\frac{\|Q_\tau^\star - Q_\tau^{(0)}\|_\infty \gamma}{\epsilon}\right)$$

Global linear convergence of entropy-regularized NPG
at a rate independent of $|\mathcal{S}|$, $|\mathcal{A}|$!

Comparisons with entropy-regularized PG



(Mei et al., 2020) showed entropy-regularized PG achieves

$$V_\tau^\star(\rho) - V_\tau^{(t)}(\rho) \le \left(V_\tau^\star(\rho) - V_\tau^{(0)}(\rho)\right)$$

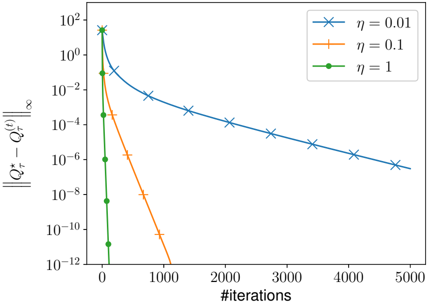
$$\cdot \exp\left(-\frac{(1-\gamma)^4 t}{(8/\tau + 4 + 8\log|\mathcal{A}|)|\mathcal{S}|} \left\| \frac{d_\rho^{\pi_\tau^\star}}{\rho} \right\|_\infty^{-1} \min_s \rho(s) \underbrace{\left(\inf_{0\le k\le t-1} \min_{s,a} \pi^{(k)}(a|s) \right)^2}_{\text{can be exponential in } |\mathcal{S}| \text{ and } \frac{1}{1-\gamma}} \right)$$

Much faster convergence of entropy-regularized NPG
at a **dimension-free** rate!

Comparison with unregularized NPG

Regularized NPG

$\tau = 0.001$

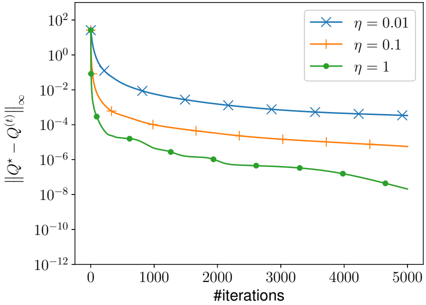


Linear rate: $\frac{1}{\eta\tau}\log\left(\frac{1}{\epsilon}\right)$

Ours

Vanilla NPG

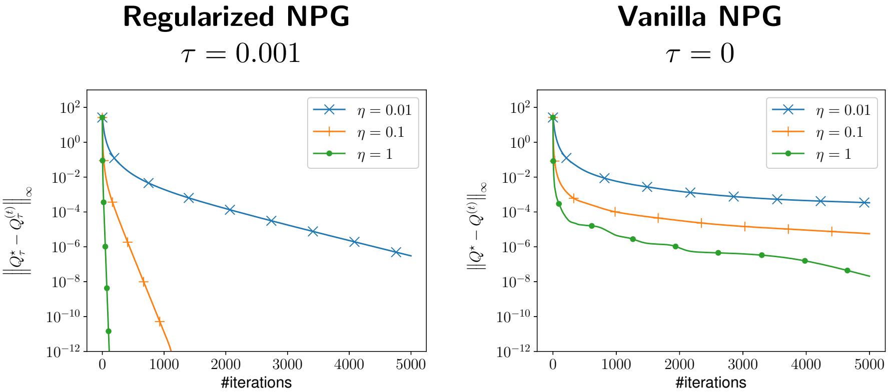
$\tau = 0$



Sublinear rate: $\frac{1}{\min\{\eta,(1-\gamma)^2\}\epsilon}$

(Agarwal et al. 2019)

Comparison with unregularized NPG



Linear rate: $\frac{1}{\eta\tau}\log\left(\frac{1}{\epsilon}\right)$

Ours

Sublinear rate: $\frac{1}{\min\{\eta,(1-\gamma)^2\}\epsilon}$

(Agarwal et al. 2019)

Entropy regularization enables fast convergence!

So far, we assume complete knowledge of Q-function for each π_t...

Entropy-regularized NPG with inexact gradients

Inexact oracle: inexact evaluation of $Q_\tau^{(t)}$, which returns $\widehat{Q}_\tau^{(t)}$ s.t.

$$\left\|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\right\|_\infty \le \delta,$$

e.g. using sample-based estimators

Entropy-regularized NPG with inexact gradients

Inexact oracle: inexact evaluation of $Q_\tau^{(t)}$, which returns $\widehat{Q}_\tau^{(t)}$ s.t.

$$\left\| \widehat{Q}_\tau^{(t)} - Q_\tau^{(t)} \right\|_\infty \leq \delta,$$

e.g. using sample-based estimators

Inexact entropy-regularized NPG:

$$\pi^{(t+1)}(a|s) \;\propto\; \left(\pi^{(t)}(a|s)\right)^{1-\frac{\eta\tau}{1-\gamma}} \exp\left(\frac{\eta \widehat{Q}_\tau^{(t)}(s,a)}{1-\gamma}\right)$$

Entropy-regularized NPG with inexact gradients

Inexact oracle: inexact evaluation of $Q_\tau^{(t)}$, which returns $\widehat{Q}_\tau^{(t)}$ s.t.

$$\left\|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\right\|_\infty \le \delta,$$

e.g. using sample-based estimators

Inexact entropy-regularized NPG:

$$\pi^{(t+1)}(a|s) \;\propto\; \left(\pi^{(t)}(a|s)\right)^{1-\frac{\eta\tau}{1-\gamma}} \exp\left(\frac{\eta \widehat{Q}_\tau^{(t)}(s,a)}{1-\gamma}\right)$$

Question: stability vis-à-vis inexact gradient evaluation?

Linear convergence with inexact gradients

$$\big\|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\big\|_\infty \le \delta$$

Theorem (Cen, Cheng, Chen, Wei, Chi '22)

For any stepsize $0 < \eta \le (1-\gamma)/\tau$, entropy-regularized NPG attains

$$\big\|Q_\tau^\star - Q_\tau^{(t+1)}\big\|_\infty \le \gamma(1-\eta\tau)^t C_1 + C_2$$

- $C_1 = \|Q_\tau^\star - Q_\tau^{(0)}\|_\infty + 2\tau\Big(1 - \dfrac{\eta\tau}{1-\gamma}\Big)\|\log\pi_\tau^\star - \log\pi^{(0)}\|_\infty$
- $C_2 = \dfrac{2\gamma\big(1+\frac{\gamma}{\eta\tau}\big)}{(1-\gamma)^2}\delta$: error floor

- converges linearly at the same rate until an error floor is hit

Linear convergence with inexact gradients

$$\big\|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\big\|_\infty \le \delta$$

Theorem (Cen, Cheng, Chen, Wei, Chi '22)

For any stepsize $0 < \eta \le (1-\gamma)/\tau$, *entropy-regularized NPG attains*

$$\big\|Q_\tau^\star - Q_\tau^{(t+1)}\big\|_\infty \le \gamma(1-\eta\tau)^t C_1 + C_2$$

- $C_1 = \|Q_\tau^\star - Q_\tau^{(0)}\|_\infty + 2\tau\Big(1 - \dfrac{\eta\tau}{1-\gamma}\Big)\|\log\pi_\tau^\star - \log\pi^{(0)}\|_\infty$
- $C_2 = \dfrac{2\gamma\big(1+\frac{\gamma}{\eta\tau}\big)}{(1-\gamma)^2}\delta$: error floor

- converges linearly at the same rate until an error floor is hit
- sample complexity $\widetilde{O}\Big(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^8}\epsilon^2\Big)$ (sub-optimal)

Returning to the original MDP?

How to employ entropy-regularized NPG to find an ε-optimal policy for the original (unregularized) MDP?

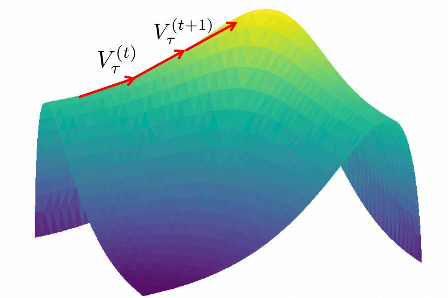
Returning to the original MDP?

How to employ entropy-regularized NPG to find an ε-optimal policy for the original (unregularized) MDP?

- suffices to find an $\frac{\varepsilon}{2}$-optimal policy of regularized MDP w/ regularization parameter $\tau = \frac{(1-\gamma)\varepsilon}{4 \log |\mathcal{A}|}$
- iteration complexity is the same as before (up to log factor)

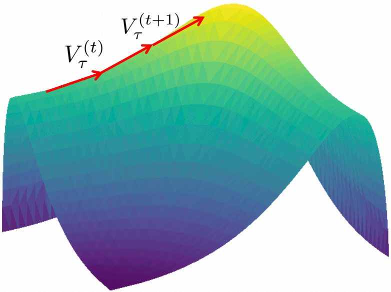
A warm-up analysis when $\eta = \frac{1-\gamma}{\tau}$

A key lemma: monotonic performance improvement



$$V_\tau^{(t+1)}(\rho) - V_\tau^{(t)}(\rho) = \mathop{\mathbb{E}}_{s\sim d_\rho^{(t+1)}}\left[\left(\frac{1}{\eta} - \frac{\tau}{1-\gamma}\right)\underbrace{\mathsf{KL}\Big(\pi^{(t+1)}(\cdot|s)\,\big\|\,\pi^{(t)}(\cdot|s)\Big)}_{\text{KL divergence}} + \frac{1}{\eta}\underbrace{\mathsf{KL}\left(\pi^{(t)}(\cdot|s)\,\big\|\,\pi^{(t+1)}(\cdot|s)\right)}_{\text{KL divergence}}\right]$$

A key lemma: monotonic performance improvement



$$V_\tau^{(t+1)}(\rho) - V_\tau^{(t)}(\rho) = \mathop{\mathbb{E}}_{s\sim d_\rho^{(t+1)}}\Bigg[\left(\frac{1}{\eta} - \frac{\tau}{1-\gamma}\right)\underbrace{\mathsf{KL}\Big(\pi^{(t+1)}(\cdot|s)\,\big\|\,\pi^{(t)}(\cdot|s)\Big)}_{\text{KL divergence}}$$

$$+ \frac{1}{\eta}\underbrace{\mathsf{KL}\left(\pi^{(t)}(\cdot|s)\,\big\|\,\pi^{(t+1)}(\cdot|s)\right)}_{\text{KL divergence}}\Bigg]$$

$$\geq 0 \qquad \left(\text{if } 0 < \eta \leq \frac{1-\gamma}{\tau}\right)$$

A key operator: soft Bellman operator

Soft Bellman operator

$$\mathcal{T}_\tau(Q)(s,a) := \underbrace{r(s,a)}_{\text{immediate reward}} + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\left[\max_{\pi(\cdot|s')} \mathop{\mathbb{E}}_{a'\sim\pi(\cdot|s')}\Big[\underbrace{Q(s',a')}_{\text{next state's value}} - \underbrace{\tau \log \pi(a'|s')}_{\text{entropy}}\Big]\right],$$

A key operator: soft Bellman operator

Soft Bellman operator

$$\mathcal{T}_\tau(Q)(s,a) := \underbrace{r(s,a)}_{\text{immediate reward}} + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left[\max_{\pi(\cdot|s')} \mathop{\mathbb{E}}_{a' \sim \pi(\cdot|s')} \Big[\underbrace{Q(s',a')}_{\text{next state's value}} - \underbrace{\tau \log \pi(a'|s')}_{\text{entropy}} \Big] \right],$$

Soft Bellman equation: $Q_\tau^\star$ is *unique* solution to

$$\mathcal{T}_\tau(Q_\tau^\star) = Q_\tau^\star$$

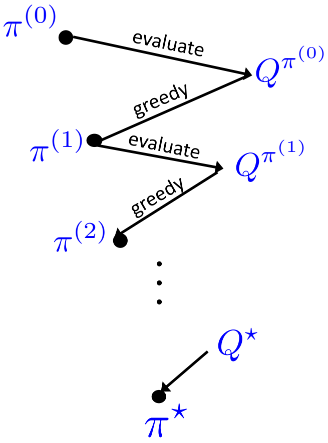
γ-contraction of soft Bellman operator:

$$\|\mathcal{T}_\tau(Q_1) - \mathcal{T}_\tau(Q_2)\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty$$

Richard Bellman

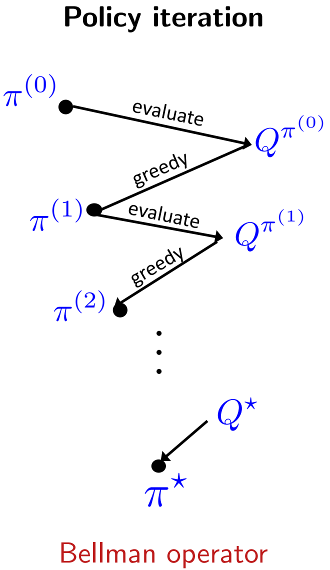
Analysis of soft policy iteration ($\eta = \frac{1-\gamma}{\tau}$)

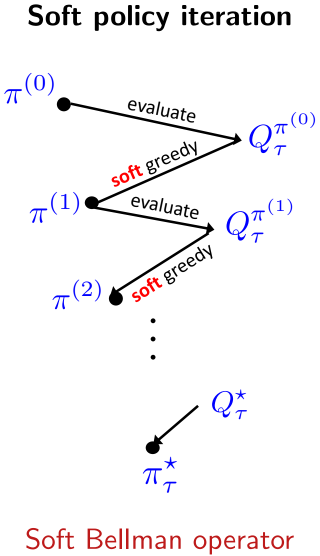
Policy iteration



Bellman operator

Analysis of soft policy iteration ($\eta = \frac{1-\gamma}{\tau}$)





A key linear system: general learning rates

Let $x_t := \begin{bmatrix} \big\|Q_\tau^\star - Q_\tau^{(t)}\big\|_\infty \\ \big\|Q_\tau^\star - \tau \log \xi^{(t)}\big\|_\infty \end{bmatrix}$ and $y := \begin{bmatrix} \big\|Q_\tau^{(0)} - \tau \log \xi^{(0)}\big\|_\infty \\ 0 \end{bmatrix}$,

where $\xi^{(t)} \propto \pi^{(t)}$ is an auxiliary sequence, then

A key linear system: general learning rates

Let $x_t := \begin{bmatrix} \|Q_\tau^\star - Q_\tau^{(t)}\|_\infty \\ \|Q_\tau^\star - \tau \log \xi^{(t)}\|_\infty \end{bmatrix}$ and $y := \begin{bmatrix} \|Q_\tau^{(0)} - \tau \log \xi^{(0)}\|_\infty \\ 0 \end{bmatrix}$,

where $\xi^{(t)} \propto \pi^{(t)}$ is an auxiliary sequence, then

$$x_{t+1} \leq A x_t + \gamma \left(1 - \frac{\eta\tau}{1-\gamma}\right)^{t+1} y,$$

where

$$A := \begin{bmatrix} \gamma \\ 1 \end{bmatrix} \cdot \begin{bmatrix} \frac{\eta\tau}{1-\gamma} & 1 - \frac{\eta\tau}{1-\gamma} \end{bmatrix}$$

is a rank-1 matrix with a non-zero eigenvalue $\underbrace{1 - \eta\tau}_{\text{contraction rate!}}$.

Beyond entropy regularization

Leverage regularization to promote structural properties of the learned policy.

cost-sensitive RL

weighted 1-norm

sparse exploration

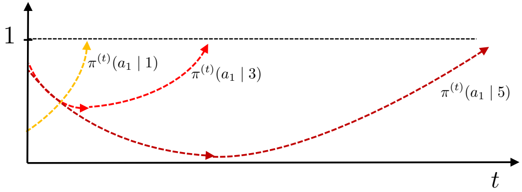
Tsallis entropy

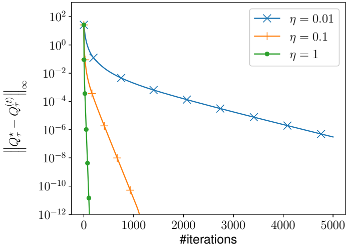
constrained and safe RL

log-barrier

For further details, see: (Lan, PMD 2021) and (Zhan et al, GPMD 2021)

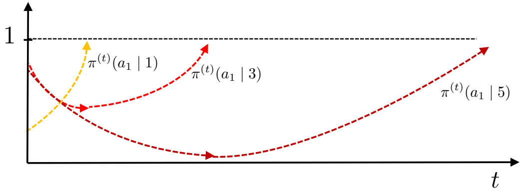
Summary of this part

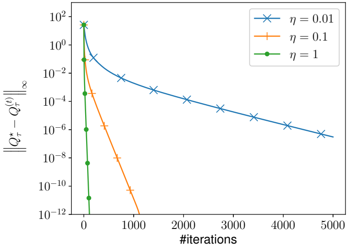




- Softmax policy gradient can take exponential time to converge
- Entropy regularization & natural gradients help!

Summary of this part





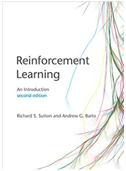
- Softmax policy gradient can take exponential time to converge
- Entropy regularization & natural gradients help!

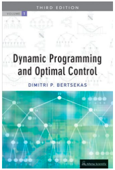
Future directions:

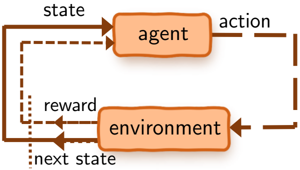
- optimal sample complexity bound
- function approximation

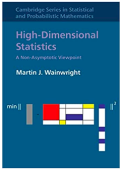
Concluding Remarks

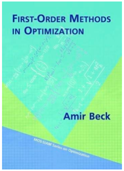
Concluding remarks











Understanding non-asymptotic performances of RL algorithms is a fruitful playground!

Promising directions:

- function approximation
- multi-agent/federated RL
- hybrid RL
- many more...

Thank you for your attention! https://yutingwei.github.io/